2001

BUILDING DSS USING KNOWLEDGE DISCOVERY IN DATABASE APPLIED TO ADMISSION & REGISTRATION FUNCTIONS

EL-RAGAL, AHMED ABDEL HAMEED HASSAN

BUILDING DSS USING KNOWLEDGE DISCOVERY IN DATABASE APPLIED TO ADMISSION & REGISTRATION FUNCTIONS

by

AHMED ABDEL HAMEED HASSAN EL-RAGAL

A thesis submitted to the University of Plymouth Business School
In partial fulfillment for the degree of

DOCTOR OF PHILOSOPHY

IN INFORMATION SYSTEMS

The University of Plymouth Business School

PH.D.

November 2001

BUILDING DSS USING KNOWLEDGE DISCOVERY IN DATABASE APPLIED TO ADMISSION & REGISTRATION FUNCTIONS

A. A.H. H. EL-RAGAL

PH.D.

November 2001

PLYMOUTH

Dedication

To the innocent souls of my parents,

My lovely wife and daughter,

And special dedication to my supervisor; Terry Mangles

AHMED ABDEL HAMEED HASSAN EL-RAGAL

BUILDING DSS USING KNOWLEDGE DISCOVERY IN DATABASE APPLIED TO ADMISSION & REGISTRATION FUNCTIONS

ABSTRACT

This research investigates the practical issues surrounding the development and implementation of Decision Support Systems (DSS). The research describes the traditional development approaches analyzing their drawbacks and introduces a new DSS development methodology. The proposed DSS methodology is based upon four modules; needs' analysis, data warehouse (DW), knowledge discovery in database (KDD), and a DSS module.

The proposed DSS methodology is applied to and evaluated using the admission and registration functions in Egyptian Universities. The research investigates the organizational requirements that are required to underpin these functions in Egyptian Universities. These requirements have been identified following an in-depth survey of the recruitment process in the Egyptian Universities. This survey employed a multi-part admission and registration DSS questionnaire (ARDSSQ) to identify the required data sources together with the likely users and their information needs. The questionnaire was sent to senior managers within the Egyptian Universities (both private and government) with responsibility for student recruitment, in particular admission and registration.

Further, access to a large database has allowed the evaluation of the practical suitability of using a data warehouse structure and knowledge management tools within the decision making framework. 1600 students' records have been analyzed to explore the KDD process, and another 2000 records have been used to build and test the data mining techniques within the KDD process.

Moreover, the research has analyzed the key characteristics of data warehouses and explored the advantages and disadvantages of such data structures. This evaluation has been used to build a data warehouse for the Egyptian Universities that handle their admission and registration related archival data. The decision makers' potential benefits of the data warehouse within the student recruitment process will be explored.

The design of the proposed admission and registration DSS (ARDSS) will be developed and tested using Cool: Gen (5.0) CASE tools by Computer Associates (CA), connected to a MS-SQL Server (6.5), in a Windows NT (4.0) environment. Crystal Reports (4.6) by Seagate will be used as a report generation tool. CLUSTAN Graphics (5.0) by CLUSTAN software will also be used as a clustering package.

Finally, the contribution of this research is found in the following areas:

- A new DSS development methodology;
- The development and validation of a new research questionnaire (i.e. ARDSSQ);
- The development of the admission and registration data warehouse;
- The evaluation and use of cluster analysis proximities and techniques in the KDD process to find knowledge in the students' records;
- And the development of the ARDSS software that encompasses the advantages of the KDD and DW and submitting these advantages to the senior admission and registration managers in the Egyptian Universities.

The ARDSS software could be adjusted for usage in different countries for the same purpose, it is also scalable to handle new decision situations and can be integrated with other systems.

Table of contents

List of tables

Chapter five: The blend of DSS, DW, and KDD

Chapter six: Extracting users' requirements

Chapter seven: The proposed Admission and Registration DSS

List of figures

Chapter five: The blend of DSS, DW, and KDD

Chapter seven: The blend of DSS, DW, and KDD

List of abbreviations

Abbreviation	*Full word*
1:1	One-to-One
1:M	One-to-Many
11 MPTCBIS	1-1 The managers' perspectives towards computers and their current Admission and Registration information systems
12 FEIS	1-2 Features of these information systems
13 FUIS	1-3 Functions of these information systems
21 MPTICBIS	2-1 The managers' perspectives towards the role of computers and the ideal Admission and Registration information system
22 DSSDE	2-2 The decisions that this DSS is expected to take
23 DSSFU	2-3 DSS functions
24 DSSCH	2-4 DSS characteristics
4GL	Fourth Generation Language
AASTMT	Arab Academy for Science and Technology and Maritime Transport
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ARDSS	Admission and Registration Decision Support System
ARDSSQ	Admission and Registration Decision Support Systems Questionnaire
BLOBS	Binary Large Objects
CAD	Computer Aided Design
CAM	Computer Aided Manufacturing
CART	Classification and Regression Techniques
CASE	Computer Aided Software Engineering
CBIS	Computer Based Information Systems
CIM	Computer Integrated Manufacturing
CSF	Critical Success Factors
CU	Constant Update
DB	Database
DBA	Database Administrator
DBMS	Database Management Systems
DFD	Data Flow Diagram
DP	Data Processing

DSS	Decision Support Systems
DW	Data Warehouse
EDA	Exploratory Data Analysis
EDP	Electronic Data Store
EDW	Enterprise Data Warehouse
EIS	Executive Information Systems
ES	Expert Systems
ESS	Executive Support Systems
E.S.S.	Error Sum of Squares
FAQ	Frequently Asked Questions
FN	Furthest Neighbour
GA	Genetic Algorithms
GDSS	Group Decision Support Systems
GEIS	Global Executive Information Systems
GIS	Geographic Information Systems
GUI	Graphical User Interface
I-CASE	Integrated Computer Aided Software Engineering
IE	Information Engineering
IS	Information Systems
IT	Information Technology
IUV	Independently Update Views
JAD	Joint Application Development
KBDSS	Knowledge Based Decision Support Systems
KDD	Knowledge Discovery in Database
KMS	Knowledge Management Systems
LAN	Local Area Network
LINX	London Internet Exchange
M:1	Many-to-One
M:N	Many-to-Many
MBMS	Model Base Management Subsystems
MDDBMS	Multi-Dimensional Database Management Systems
MDDM	Multi-Dimensional Data Modelling
MIS	Management Support Systems

MMDBMS	Multi-Media Database Management Systems
NN	Nearest Neighbour
OAS	Office Automation Systems
ODS	Operational Data Store
OLAP	On-line Analytical Processing
OLTP	On-line Transaction Processing
OODBMS	Object Oriented Database Management Systems
OOP	Object Oriented Programming
PCR	Parent Child Relationship
PK	Primary Key
RDBMS	Relational Database Management Systems
RI	Referential Integrity
RT	Read Transaction
SDLC	Systems Development Life Cycle
SIS	Strategic Information Systems
SQL	Structured Query Language
TPS	Transactions Processing Systems
UIMS	User Interface Management Subsystems
VU	Variable Update
WAN	Wide Area Network
WWW	World Wide Web

ACKNOWLEDGEMENTS

This work is attributable to many efforts without which nothing could have been available by now. I would like firstly to thank God for giving me the will to complete this work.

I would also like to thank my employer, the Arab Academy for Science and Technology and Maritime Transport in Alexandria-Egypt, for their financial support.

I would also like to thank Ms. Suzanne Martin, the Technical Support Consultant in Computer Associates-London, for spending her valuable time to review my DSS model.
Also thanks to Mr.Tarek Nofal, the Technical Support Manager in Systems Integrators-Cairo, for spending his time to review my DSS model.

Moreover, I really appreciate everyone's time and effort that were spent with me during the questionnaire development and testing stages; Paul Brand (University of Plymouth Business School), David Flemming (University of Plymouth MIS department), Ahmed Gomaa (Nova South Eastern University-Florida), Mahmoud Youssef (Rutgers University-New Jersey), Hossam Youness (Qatar), Dr. Bahgat A.Maksoud (Assiut University), Dr. A. Fattah Ghobashy (Suez Canal University), Dr. Wafaa Hanafy (Menoufia University), Mrs. Mervat Hanafy (Tanta University), and Dr. Nagy Elsemelawy (AASTMT).

Special thanks go to the Business team and the Inter Library Loans at our fabulous University of Plymouth Library.

Finally thanks to the University of Plymouth Business School for the facilities and the continuous support.

My greatest appreciation for my family and friends for their encouragement that pushed me forward.

AUTHOR'S DECLARATION

At no time during the registration for the degree of Doctor of Philosophy has the author registered for any other University award.

The study was financed with the financial support from the Arab Academy for Science and Technology and Maritime Transport, Alexandria-Egypt.

The following activities were undertaken in connection with the programme of study:

- Attendance and participation in research seminars, during which research work was presented.
- Presentation of papers from the thesis in conferences are as follows:

1. El-Ragal, A. and Mangles, T., "Knowledge Discovery in Database Techniques Applied to Students Recruitment Systems in Universities", **The 15th IAIM 2000 Conference**, in Brisbane, Australia, 8-10 December 2000, pp. 7-20.
2. El-Ragal, A. and Mangles, T., "Developing a Star Schema Structures- A Practical Study", **The 4th SAIS 2001 Conference**, in Savannah-GA, USA, 2-3 March 2001, pp. 116-136.
3. El-Ragal, A., Mangles, T., and Chaston, I., "Developing a Star Schema Structures- A Practical Study applied to the Egyptian Universities", **The BIT World 2001 Conference**, in Cairo, Egypt, 4-6 June 2001, pp. 128.
4. Mangles, T., and El-Ragal, A., "Developing a new Decision Support System for University Student Recruitment", to be presented at **The ICEB Conference**, in Hong Kong, 19-21 December 2001.

Signed: .....

Date: 20/11/2001

Chapter One

Introduction

1-1 Introduction

The type of information required by managers is directly related to the level of management and the amount of structure in the decision situations they face. Decisions at the operational level tend to be more structured, those at the tactical level more semi-structured, and those at the strategic level more unstructured. Structured decisions involve situations where the procedures to follow to reach a decision can be specified in advance. Inventory reorder decisions faced by businesses are a typical example (O'Brien, 1996). Unstructured decisions involve decision situations where it is not possible to specify in advance most of the decision procedures to follow. At most, many decision situations are semi-structured. That is, some decision procedures can be pre-specified, but not enough to lead to a definite recommended decision. For example, decisions involved in starting a new line of products or making a major change to employee benefits would probably range from unstructured to semi-structured.

Therefore, information systems must be designed to produce a variety of information products to meet the changing decision needs of managers at different levels of an organization. For example, the strategic management level requires more summarized, ad hoc, unscheduled reports, forecasts, and external intelligence to support its more unstructured planning and policy making responsibilities. The operational management level, on the other hand, may require more regular internal reports emphasizing detailed current and historical comparisons that support its more structured control of day-to-day operations. Thus, we can generalize that higher levels of management require more ad hoc, unscheduled, infrequent summaries, with a wide, external, and forward-looking scope. On the other hand, lower levels of management require more pre-specified, frequently scheduled, and detailed information, with a more narrow, internal, and historical focus (O'Brien, 1996).

However, managers use different information systems for each class of decisions (Long and Long, 2001; Turban, 1993). For structured decisions Transaction Processing Systems (TPS) and Management Information systems (MIS) are usually used. For the unstructured decisions, Decision Support Systems (DSS) Expert Systems (ES) and Artificial Neural Networks (ANN) are used. Executive Information Systems (EIS) are special types of information systems that support unstructured decisions (Long and Long, 2001; Turban, 1993; Long, 1989).

Adam, et al. (1997: 2) said, "The DSS area is probably the most widely researched in the information systems field and is one that continues to be a focus for information systems researchers."

1-2 DSS

The concepts involved in DSS were first articulated in 70's by Scott-Morton. According to them, DSS are interactive computer-based systems, which help decision-makers utilize data and models to solve unstructured problems.

Little (1970) defines DSS as model-based set of procedures for processing data and judgments to assist a manager in his decision making. He argues that such systems must be:

1. Simple;
2. Robust;
3. Easy to control;
4. Adaptive;
5. Complete on important issues;
6. Easy to communicate with.

Keen and Scott-Morton (1978) provided another classical definition of DSS which said that Decision Support Systems couple the intellectual resources of individuals with the capabilities of the computer to improve the decisions' quality. DSS are computer-based support systems for the decision maskers who deal with semi-structured problems.

Keen (1980) applied the term DSS to situations where a "*final*" system can only be developed through an adaptive process of learning and "*evolution*".

Bonczek, et al. (1980) defined DSS as computer-based systems consisting of three interacting components:

1. A language system;
2. A knowledge system;
3. A problem processing system - the link between the other two components.

A study by Stevens (1991) evaluated the use of DSS in the banking industry in England. The results of Stevens's study showed that the use of DSS has evolved from being a simple system that only accessed the corporate data and reported it, to become knowledge based systems. No details on such knowledge based systems were given i.e. components, technologies used,...etc.

Klein and Methlie (1995) defined DSS as computer programs that provide information in a specific application domain by means of analytical decision models and access to databases in order to support decision makers effectively in semi-structured and unstructured decisions. They also emphasized that end users of a specific DSS are not always known during the development but what is well known is the problem. Klein and Methlie discussed the data sources of the DSS, ignored completely the use of historical or

archival data sources. Their work was focused on the types of problems which the DSS can handle, the data models used, the users of the DSS, and the objectives of having DSS.

A study by Barron and Saharia (1995) suggested a storage structure that will contain the results of any statistical queries which are accessed frequently by the DSS users. They claimed that storing the results from these queries would enhance the DSS performance. The study did not handle the problems associated with storing results from statistical queries whilst the data sources were being updated. That means that the stored results do not necessarily reflect the most up-to-date data values. This problem is handled by the use of summary tables in data warehousing. However Barron and Saharia did not mention any warehousing components.

A study by Hawk and Bariff (1995) examined the organisational strategies for supporting DSS (e.g. traditional application development groups, DSS groups within end-user services). They said that DSS are increasing exponentially in organisations, but little is known about how these organisations support their DSS users. They emphasized that support practices affect the DSS groups' ability to provide support services. They interviewed managers of twenty three DSS groups (they contacted fifty seven organisations, twenty three accepted to participate). The results showed that the support characteristics tend to vary when comparing groups in different locations, suggesting that certain strengths and limitations could be associated with organisational support strategies. The factors that affect the DSS support services are; staff-to-user ratio, staff background, use of evolutionary development methodology, and functional area of the users.

A study by Ramirez, et al. (1996) focused on the data structure on which the results obtained by what-if based DSS are stored. Ramirez, et al. called this data structure independently updated views (IUVs). IUVs are used to store derived data from the DSS

without losing consistency with the original data. They used SQL to create and maintain these IUVs. Ramirez, et al. study is redundant, as the IUVs they proposed are basically summary tables that are used to store frequently accesses queries, and it would have been easier if they had employed a data warehouse component to resolve the update problems and this would also allow the system to handle larger number of records than the proposed IUVs.

A study by Barr and Sharda (1997) on the effectiveness of DSS revealed that the research efforts on the DSS outcomes, especially those that take the longitudinal effect, are very few. Barr and Sharda suggest that the effect of DSS on decision outcomes develops over time. They claimed that the DSS effect is due to two factors; DSS development and the reliance effect (i.e. the decision maker defers the decision because the computer will take it later). Their research concluded that both factors affect decision makers, that is, reliance has a short-term effect whilst DSS development has a long-term effect. Finally, they concluded that DSS aided decision makers outperform non-DSS decision maker in organisations.

Another study by Bhargava, et al. (1997) found that there are many ways and various software systems that an organisation can use to resolve a decision problem. Bhargava, et al. claimed that database management systems and the other decision technologies (i.e. any kind of computational procedures that have the ability to support decision making) are little used in real world applications because the market that distributes them is not professional enough. Bhargava, et al. suggested an electronic market for decision technologies (Electronic market is a market where the enabling medium for transactions between consumers, providers, and services is an information network. E.g. the World Wide Web). They introduced their DecisionNet which performs functions (user accounts,

billing, setting up interfaces) that would otherwise have been needed to be developed for each consumer and/or technology.

A study by Ahn and Ezawa (1997) focused on the knowledge-based decision support systems (KBDSS). They claimed that many KBDSS have been developed, however few systems address the use of knowledge in the decision problem. Their study suggested a KBDSS for the telemarketing industry that predicts the probability of a customer disconnecting the telephone line based on previous promotions and customer history. The prediction model used a Bayesian network model linked with an influence diagram. Ahn and Ezawa study was based on the customer's history (i.e. transactions), however their study did not include any data warehouse components.

Adam, et al. (1998) study was aimed at the classification of DSS usage in organisations. Their study followed the framework introduced by Gorry and Scott-Morton. However, Adam, et al. suggested that DSS usage is also based on two dimensions; DSS spread and DSS complexity. Their study included seventeen organisations and the research results are summarized in the following table (1-1):

Problem type	**Management level**		
	Low/Operational	**Middle/Control**	**Top/Strategic**
Structured	-Accounts receivable -Order entry	-Short term forecasting	-Factory location
Semi-structured	-Production scheduling	-Budget analysis	-Mergers -Acquisitions
Unstructured	-Cash management	-Sales and production	-New product line

Table (1-1). DSS usage across organisations[1].

[1] Adapted from Adam, et al. (1998).

Lee, et al. (1999) focused on groups of people taking decisions together, using what is known as group decision support systems (GDSS). GDSS allow a variety of specialists to be assembled whereby each of them is contributing to the solution using his expertise. GDSS could stimulate creative thinking and allow people from different departments to take the decision together. The disadvantage of GDSS is the possible conflict of the people contributing to the solution because each has a departmental view. However, when all of them share the organisational goals this could encourage them to be committed to the organisation rather than individual departments. Lee, et al. concentrated on the type of models and decisions taken rather than the components and technologies that should be used to implement GDSS.

Long and Long (2001) defined DSS as interactive information systems that rely on an integrated set of user-friendly decision support tools to produce information to support management in the decision making process. Long and Long claimed that for simple and structured decisions managers use their own experience, whilst for complex and unstructured decisions manager use DSS to close the gap between what they can do and what the problem requires. They explained that a data warehouse could be a data management component of DSS. Long and Long also raised the importance of having data mining techniques as analytical tools provided with the DSS to find hidden knowledge in the data warehouse and provide that knowledge to the decision maker enabling better quality decisions to be made. In their discussion of a data warehouse they claimed that the warehouse is a relational database, which is not always the case. The data warehouse is either relational model of star schema structured (Refer to chapter three, section 3-9 for more details).

1-3 Data Warehousing

Not all the databases at an organization have the same classification. For example, the running applications have certain database design requirements and for this reason, these types of databases are known as *operational databases*. They are not designed to respond to spontaneous queries, however, they are optimized for carrying high speed and large numbers of users. Another type of database that can be found in organizations is the data warehouse (DW). *This is designed for strategic decision support, and is largely built up from the operational databases*. The DW can contain a large amount of data and millions of data records. Smaller, local data warehouses are called *data marts* (Adriaans and Zantinge, 1996).

A data warehouse is the means for strategic data usage (Berson, 1996). In other words, a data warehouse is a blend of technologies aimed at effective integration of operational databases into an environment that enables the strategic use of data.

Widom (1995) said that data warehousing was introduced to build a logically centralized data repository to fulfill the requirements of strategic data usage and prevent local systems from DW users competition.

The purpose of a DW is to establish a data repository that makes the operational data accessible in a form that is readily acceptable for the analysis. Only the data required to meet the executives' needs are taken from the operational environment.

DW can be viewed as an information system (IS) foundation that has the following characteristics (Berson, 1996):

1. It is used intensively for READ type transactions;
2. It includes a large volume of records in a few number of tables;
3. Each query is processed in large data sets using multi-table joins;

4. It contains current and historical data;
5. It is periodically updated;
6. It supports a small number of users;
7. It is a database designed for analytical tasks;
8. It uses data from different databases, from various applications.

Taha, et al. (1997) study gave some examples of the users who are considered the DW primary candidates:

1. Users who need certain data presented in a special format of summarization and aggregation;
2. Users who deal with historical data;
3. Users who need to reply to the frequently asked queries-FAQ;
4. Users who need a continuous accessing of certain data.

Looking to the previous list with more insight reveals the fact that these are not operational issues e.g. sales ordering processing, but are aimed at providing information for the organisations' decision makers and executives who are the primary candidates to use the DW.

A study by Sørensen and Alnor (1999) focused on the creation of a data warehouse using SQL Server. The study did not include how to use front-end tools to utilize the DW. The study also did not mention anything about the DW indexing strategy.

Long and Long (2001) focused on the data warehouse component as part of the DSS. They claimed that adding a DW component to the DSS would enhance the quality of decisions by providing a broad range of data sources (archival and external) to the decision makers.

Wixom and Watson (2001) studied the factors that affect the data warehouse success. Their study showed that resources, user participation, highly skilled project team members increase the likelihood that warehousing projects will end on time within budget constraints and with proper functionality. Wixom and Watson also found out that the implementation success with organizational and project issues will in turn affect the system quality of the data warehouse.

1-4 Knowledge Discovery and data mining

Knowledge Discovery in Databases (KDD) means a process of nontrivial extraction of implicit, previously unknown and potentially useful information (Fayyad, et al., 1996). This discovered knowledge could be of great value in many areas, foremost among which is decision making (Chen, et al., 1996). Moulet and Kodratoff (1995) defined the KDD as a three-step process that involves data selection, data cleaning, and finally knowledge interpretation.

KDD is not a separate body of knowledge that stands for its own. KDD is a multi-discipline branch of science; databases, statistics, visualization techniques, machine learning, and expert systems. They all contributed to the KDD process.

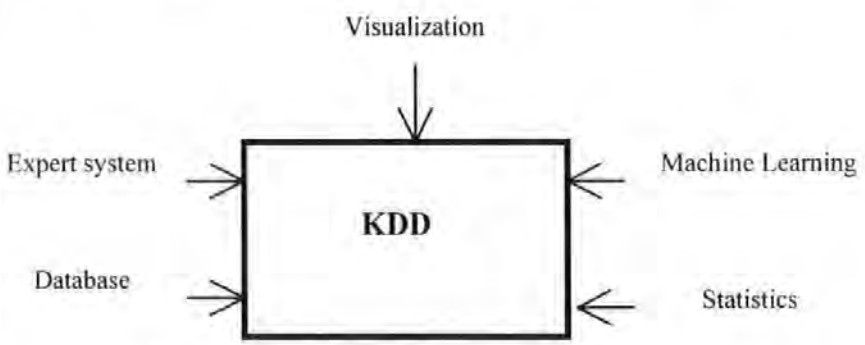


Figure (1-1). KDD is a multi-disciplinary field[2].

On the other hand, data mining gives organisations the tools to navigate through large amounts of data to find trends and patterns which can guide the strategic decision making

[2] Adapted from (Adriaans and Zantinge, 1996).

process. Data mining is not specific to any particular industry- it requires intelligent technology and the willingness to explore the possibility of hidden knowledge that resides in the data. The results of Taha, et al. (1997) study suggest that data warehousing is not a mandatory component in establishing the data mining environment, that is, data mining could be established without using a data warehouse.

There is confusion about the exact meaning of the two terms '*data mining*' and '*KDD*', with many authors regarding them as synonymous (Ganti, et al., 1999; Adriaans and Zantinge, 1996). At the first international KDD conference in Montreal in 1995, it was proposed that the term 'KDD' would be employed to describe the whole process of extraction of knowledge from data. In this context, *knowledge means understanding the relationships and patterns between data elements.* It was further proposed that the term data mining should be used exclusively for the discovery stage of the KDD process. For the purpose of this research, data mining techniques will be viewed as a step in the KDD process.

Silberschatz and Tuzhilin (1995) study on the interestingness in knowledge discovery results revealed that the more actionable the discovered patterns are, the more interested the user is and further the user will apply those patterns and update his experience accordingly. The less interesting the patterns the less actionability and it is more likely that the user will discard the discovered pattern.

Syan, et al. (1996) gave an introduction to data mining from a database perspective. Their results showed that data mining is a fast expanding field with many new research results. Based on diversity of data mining methods and rich functionalities of data mining investigated so far, many of which have been used successfully for mining knowledge in large databases.

A different research study (Silberschatz and Tuzhilin, 1996) discussed the degree of user involvement in the discovery process. The study outlines a spectrum of degrees of user-involvement and presents the data monitoring and discovery-triggering approaches that provides a balanced "*division of labor*" between the KDD application development and the discovery engine.

Clifton and Marks (1996) examined the security and privacy implications of data mining. They found that data mining technology provides a whole new way of exploring large databases. They also found that there is a trade off between restricting the access to data and the advantages of database and network technologies in the ability to share data.

Grossman, et al. (1996) have taken the statistical approach to dealing with data mining techniques. This approach was adopted by a number of researchers at the early data mining stages.

Berson (1996) differentiated between the operational data that is used in the traditional information systems, and informational data that is used in the information systems built around the mining techniques.

A study by McSherry (1997) on knowledge discovery by inspection revealed that simple rules can be discovered by visual inspection of frequency tables.

Liu, et al. (1999) study investigated the interestingness of the patterns resulting from the KDD process. They said that it is very easy to discover a huge number of patterns from any database, however, most of these patterns are useless and uninteresting to the users. They suggested that to prevent the users from being overwhelmed by the large number of patterns, techniques are required to rank them according to their interestingness.

Although data mining is a new strategic tool in the executive manager's hand, the strategic value of data mining is time sensitive, that is the output of data mining changes over time due to changes in the data sources.

In order to conduct effective data mining, one needs first to know what kind of features and requirements that should be available for this mining process. Chen, et al. (1996) identified these as:

1. *Handling different types of data.* Because there are many kinds of databases used in different applications, one might expect that a knowledge discovery system should be able perform data mining on different kinds of data. Since most databases available are Relational Data Base Management systems (RDBMS), it is expected that data mining techniques can find knowledge easily in these RDBMS. Moreover, data mining techniques should find knowledge in other complex data types; for example structured data, hypertext, multi-media data sets, transaction data, and legacy databases;
2. *Efficiency and scalability of data mining algorithms.* To extract information from a huge amount of data in databases, the knowledge discovery algorithm must be efficient and scaleable to large databases;
3. *Meaningful data mining results.* The output of the data mining process should be meaningful to the decision makers;
4. *Testing the data mining results.* The consistency of knowledge obtained from the data mining process can be examined by using different analysis techniques on the same data set;
5. *Mining information from different sources of data.* Mining knowledge from different formatted and unformatted data sets is a very rich technique that may enhance the results and increase the confidence level in the data mining results. Examples of sources of unformatted data sets are found in the data stored in Wide Area Networks-

WAN, Internet, and many other external databases. However, this poses the problems of non-heterogeneous databases- refer *to point number 1 in this list*;

6. *Security and privacy versus data mining.* When data can be viewed from different angles, it may threaten other people's privacy and data security mechanisms. To resolve this point, it is important to prevent the disclosure of very sensitive information;
7. *Data cleaning.* Before in-depth analysis is carried out using data mining techniques, the data should be cleaned, i.e. be error-free, so that we can increase the confidence in the results.

Adriaans and Zantinge (1996) emphasized that the combination of data warehousing, decision support, and data mining provides an innovative and totally new approach to information management. They also introduced the concept of knowledge discovery in databases (KDD) which encompasses data mining as a step toward the fulfillment of the knowledge discovery process.

Berson (1996) assured that the the traditional DSS developments differ from what the data mining techniques can provide the DSS with. Table (1-2) indicates these differences.

Traditional DSS/EIS tools	Data Mining tools
"print out last month's course booking"	"predict the next month course booking"
"list the oldest sponsors"	"define the sponsorship criteria from the data"
"what is the average age"	"explain why X is the average age"
"using this data, tell me what are the current student behavior"	"find some new patterns for student profiles"

Table (1-2). Traditional versus Data mining tools[3]

Alavi and Leidner (2001) study was focused on knowledge management systems (KMS). They said that with the increasing trend in dealing with knowledge as a resource,

[3] Adopted from (Berson, 1996).

organisations have become more interested in KMS. As a result, IS researchers have been attracted by KMS. Alavi and Leidner claimed that the objective of KMS is to support creation, transfer, and application of knowledge in organisations. They also said that KMS are multi-faceted concepts and require various backgrounds to be available. Alavi and Leidner said that in the knowledge creation process data mining techniques could be used as one of the supporting technologies, whilst when mentioning the knowledge storage process they used the term DB and electronic bulletins rather than DW.

Dowling and Hockemeyer (2001) study was focused on the assessment of knowledge. Their study results revealed that teachers' can assess a student's knowledge in a specific course by asking questions and based upon the student's answer a new question is asked. This process (i.e. asking questions) is to be stopped when the teacher has sufficient evidence on the student's state of knowledge. Dowling and Hockemeyer said that questions are adaptive and based on predetermined study objectives. A study objective is measured by a question, and objectives are related to each others. They claimed that the problem of knowledge assessment can be applied to medical diagnosis, failure analysis, and pattern recognition. Their study results were implemented in an algorithm for the assessment of knowledge written in C++ language. Dowling and Hockemeyer adaptive system could be enhanced by storing the students' answers in a DW from which useful knowledge can be extracted that will help adding new questions and better understand the students' profiles.

1-5 Data mining techniques

A lot of techniques have been used to find hidden knowledge, so data mining is not just a single technique, instead there are a number of techniques utilized during the mining process. Any technique that helps extract knowledge out of data is useful, so data mining

techniques form a heterogeneous group (Adriaans and Zantinge, 1996). There are many data mining techniques, techniques are classified according to the following factors:

- The kind of knowledge to be mined;
- The kind of database to be mined;
- The kind of techniques adopted.

Following is an illustration of some of the techniques used in data mining:

1. *Query tools*. The first step in a data mining project should always be a rough analysis of the data set using traditional query tools. By applying simple structured query language (SQL) to a data set a wealth of information can be obtained.
2. *Visualization techniques*. Visualization techniques are a very useful method of discovering patterns in data sets, and may be used at the beginning of a data mining process to get an indication for the quality of the data set and where patterns may be found. Such techniques are developing rapidly; advanced graphical techniques in virtual reality enable people to wander through artificial data spaces.
3. *On-line Analytical Processing (OLAP)*. The idea of dimensionality can be introduced here: a table with n independent attributes can be seen as an n-dimensional space. OLAP tools store their data in a special multi-dimensional format, often in computer memory. A manager can ask any question, although the data cannot be updated. There is however, an important difference between data mining and OLAP tools: OLAP tools do not learn, they create no new knowledge, and they cannot search for new solutions. Data mining is more powerful than OLAP.
4. *Association rules*. Association rules are always defined on binary attributes. These rules make it easy to describe any database. The number of possible association rules that might be found in any data source is very large. However, there is no algorithm that will automatically give us everything of interest in the database. Some rules are found to be useless (Argawal, et al., 1996).

5. *Clustering analysis*. Clustering is classifying unclassified data. Records that are close to each others share the same cluster or group. Cluster members tend to have similar or nearly similar attributes. Clustering is used for classification and prediction purposes.
6. *Decision trees*. A predictive model that can be viewed as a tree, each branch of the tree is a classification question, and the leaves of the tree are partitions of the data set within that classification.

A study by Ali and Wallace (1997) revealed that the choice of the data mining technique(s) has to be related to pre-determined goal(s). Moreover, the goal(s) of the mining technique(s) should help achieve the objective(s) of the entire KDD process in a certain application domain.

This research will apply some of these techniques to find hidden, unknown information and facts in sample records from the Admission and Registration database(s), and introduce them to the decision-maker to enable more informed decisions to be made.

1-6 Foundations of this research and the research objectives

1-6-1 Foundations of this research in literature

This research will introduce a new DSS methodology, further the research will study the existing procedures. Fundamentally the DSS definitions introduced by Garry and Scott-Morton, Little, Alter, Keen, and Bonczek are all narrow as far as they do not present a comprehensive definition that encompasses all DSS characteristics. Further, they do not deal with the new blend of technologies DSS, Data mining techniques, and Data Warehousing (Taha, et al., 1997). Even those who tried to use the knowledge as DSS component said that it is optional and independent (Turban and Aronson, 1998). Managers and decision-makers that are the primary users of DSS have an interest in past data but unfortunately it is either not available or not in a suitable form for direct use (Turban,

1993). To overcome these issues this research introduces and develops the use of knowledge discovery in database techniques, and data warehouse to the DSS development. Thus, the research study has three major components; DSS, KDD, and DW. The idea of linking these three components has its foundations in literature as follows:

1. Adriaans and Zantinge (1996: 1) said, "The combination of data warehousing, decision support, and data mining indicated an innovative approach and totally new approach to information management.";
2. Taha, et al. (1997: 77) said, "Decision Support is data access targeted to provide the information needed by the decision makers. DBs and DW with reporting and analysis tools optimized to support business decision making are the components of the DSS.";
3. Turban and Aronson (1998: 135) said, "Organisations are recognizing that their data contain a gold mine of information if they can dig it out. Consequently, they are warehousing and data mining for users to obtain information on their own and to establish relationships that were previously unknown.";
4. Gray and Watson (1999: 1) said, "Data warehouses, OLAP, and KDD are leading to new ways of performing decision support systems and creating executive information systems for data rich environments. Yet, these developments have received almost no attention from academics either in their research on in teaching.";
5. Cooper, et al. (2000: 566) said, "DW and other advances in IT are now solving the very difficult technical problems. They make it possible to organize, store, and retrieve huge volumes of information and to select critical information for a given decision. However, before organizations can realize that "grand promise" of MIS, most will have to reshape their business processes and decision making cultures to take advantage of the technology's new capabilities. This is a non-trivial transformation."

Based on the pre-stated literature foundations a decision was made to develop a new DSS methodology that will encompass these three major components (i.e. DW, KDD, and

DSS). The new development methodology will be applied to the Admission and Registration functions in Egyptian Universities.

1-6-2 Foundations of the choice of the application domain

The choice of the Admission and Registration functions in Egyptian Universities to be this research's application domain is based on the following:

1. Reports from the Ministry of Higher Education's annual meeting (different volumes from 1980: 1999) which assured the importance of having a computer based information system that is capable of automating the Admission and Registration related processes and decisions;
2. The research efforts exerted in the area of Admission and Registration functions in Egyptian Universities are few and incomplete. Examples are:
    a. Yossef (1998) study resulted in the development of a Web-based Admission and Registration information system. However, his research did not include the use of data warehousing and/or any knowledge discovery technique;
    b. A study by Fady (2000) resulted in an electronic payment Registration system. Again her research neither includes the Admission function, nor the use of data warehousing and/or knowledge discovery techniques;
    c. Another study by Makram (2000) developed a University Data Warehouse. The study did not include the use of any front-end tools (i.e. DSS or EIS);
3. The increasing number of user (i.e. Registrars, Admission Officers, Deans, and Associate Deans) complaints about their current Admission and Registration Information Systems in different Universities;
4. The amount of unsatisfied users' needs (inability of current Admission and Registration information systems to take decisions, to predict, to profile students, the inaccessibility of the historical data);

5. The need for a computer based Admission and Registration information system, that meets the users' requirements, has been raised on many occasions:

   -Annual meeting of the Arab Countries Universities' Registrars (including Egyptian Universities' Registrars);

   -Annual evaluation of some computer based Admission and Registration information systems in some Universities (e.g. Arab Academy for Science and Technology and Maritime Transport);

   -The researcher's experience as the Admission officer of the Arab Academy for Science and Technology and Maritime Transport (2 years).

1-6-3 The objectives of this research study

The focus of this research is on how to deliver information and knowledge to a specific category of business managers (i.e. Admission and Registration Managers) to understand their business problems better and hence to improve their decisions. This will be achieved throughout the investigation of the following research objectives:

1. To investigate and critically evaluate the current DSS practices: Before introducing a new DSS methodology and a new DSS definition that is able to conform with the components of the proposed methodology, the current DSS literature and practices will be reviewed and analyzed.
2. Develop a new DSS methodology: Based on the pre-stated literature in section 1-6-1, the proposed methodology will consist of the following three components:

   2-1 DSS

   2-2 Data warehousing

   2-3 Knowledge Discovery in Database Techniques. Since there are many techniques used in different contexts for achieving various goals, the techniques that will be used are the following:

   2-3-1 SQL

2-3-2 Visualization

2-3-3 Clustering Analysis

Refer to chapter five section 5-4 for the justification of the chosen techniques.

3. Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:

   3-1The managers' perspectives towards computers and their current admission and registration information systems

   3-2 Features of these information systems

   3-3 Functions of these information systems

4. Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:

   4-1The managers' perspectives towards the role of computers and the ideal admission and registration information systems

   4-2 The decisions that are expected to be taken by DSS

   4-3 DSS functions

   4-4 DSS characteristics

5. Use the proposed methodology to develop the required Admission and Registration DSS.

1-7 Research methodology

Whilst the proposed new DSS methodology has three major components (i.e. DW, KDD, and DSS), the methodology has four modules; one for each component and a module zero for the needs' analysis. Table (1-3) illustrates the methodology and its modules, the expected deliverables of each module and the tools and mechanisms used to accomplish each of which.

The Module	Deliverable	Tools and Mechanisms
Module 0:	Needs' Analysis	-Questionnaire for user requirements -Cool: Gen CASE tools for analysis
Module 1:	Building the data warehouse	-MS-SQL Server: Star schema structure -Crystal Reports: Report generation tool.
Module 2:	Knowledge from the KDD process	-MS-SQL Server -Cool:Gen CASE tools -The data mining techniques are: SQL Visualization Clustering analysis
Module 3:	Building the DSS	-Cool: Gen CASE tools for development

Table (1-3). The new DSS methodology and its modules.

In the following sections, the data used to develop and test the proposed Admission and Registration DSS (ARDSS) will be described, the population and sample will be identified, and the data analysis methods will be introduced.

1-8 Data sources

This research uses both primary and secondary data sources, as follows:

1. *Secondary data*: the research study uses secondary data sources for two reasons:
    a. A number of 1600 records extracted from an applicants' database will be used to explore the KDD process and the various data mining techniques;
    b. A further 2000 records extracted from a students' database will be stored in the data warehouse and will be used after that to generate knowledge for the

DSS. Based on the knowledge generated from the DW using the KDD techniques, the DSS soul be able to take decisions.

2. *Primary data*: the Admission and Registration DSS questionnaire (ARDSSQ) was developed, validated, and then mailed to the Egyptian Universities to describe their current Admission and Registration information systems and to identify their needs for a new DSS. The next sections illustrate the population of the questionnaire, the sample, and the response rate obtained.

1-9 The Population and sample

1-9-1 Population

The population of this questionnaire is the Egyptian Universities. These Universities are classified into two groups government and private-funded. According to the Egyptian Supreme Council of Universities statistics (1999), The UNESCO World List of Universities (2000), and The British Council Global Education and Training Information Service- Egyptian Universities (1999) there are twenty-one Egyptian Universities; eight are private and thirteen are government. Each of the twenty-one universities consists of number of colleges, schools, faculties, and/or higher institutes. The total number of colleges, schools, faculties, and/or higher institutes in the entire population is 354.

This research investigates the Admission and Registration functions taking place in the Egyptian Universities both private and government. The Admission and Registration functions in both University types are similar in many areas. That is, they both handle students' applications, map the students to the relevant academic institutions, doing course registration, grading-related jobs, graduation, class scheduling, etc.

However, there are differences in some areas between the two University types. For example, private Universities act independently when making decisions about accepting or rejecting students whilst in the government Universities these decisions are taken centrally.

Moreover, in the private Universities the Admission and Registration functions are centralized in one department for the entire University, whilst in the government Universities there is a separate Admission and Registration department in each academic institute in the University (Refer to chapter six, section 6-2 for more details).

1-9-2 The sample

As the number of Academic Institutions in Egyptian Universities is 354 distributed among 21 different Universities a decision was made to target them all and the number of Universities that will accept to participate would be used as a *sample*. All the *twenty-one* government and private universities were contacted to send their correspondence information and to notify them that they will receive the questionnaire. Only *thirteen* universities (*six private, seven government*) responded positively. The questionnaires were sent to these *thirteen* universities.

The sample size:

A number of 670 questionnaires were sent to 13 universities. Refer to table (6-4) which illustrates the number of questionnaires sent to each university.

1-9-3 The Response rate

1-9-3-1 University-wise

Out of thirteen Universities contacted, responses from twelve Universities have been received. All six Private Universities responded, Whilst six out of seven Government Universities responded. The overall response rate at the University level is **12/13 = 92.3%**.

1-9-3-2 Respondent-wise

Out of the 670 questionnaires sent, 167 returned, which gives a response rate **167/670 = 24.9%**. This response rate is adequate in this kind of research that is based on mailed questionnaires (Chan, et. al, 1998; Saunders, et. al, 1997; Teo and King, 1996).

1-10 Data analysis

Since the research study relies on both primary and secondary data sources, different data analysis methods will be used.

1-10-1 Primary data analysis techniques

The data obtained by the questionnaire will be analyzed using Chi Square (X^2) analysis and Canonical correlation analysis.

1-10-2 Secondary data analysis techniques

-The 1600 applicants' records will be analyzed using various KDD techniques (e.g. SQL, decision trees, cluster analysis, association rules, and OLAP).

-The 2000 students' records will be analyzed using the knowledge discovery in database techniques; SQL, visualization, and clustering analysis.

1-11 The proposed DSS development

The development of DSS can be completed by several approaches. These are (Turban and Aronson, 1998):

1. The use of a general-purpose programming language, such as COBOL or PASCAL. This approach was widely used in the 80's, but not much used in the 90's;
2. The use of a fourth-generation language '4GL', such as spreadsheets. This is used where the problem is relatively simple and we want to accelerate the development process;
3. The use of DSS integrated development tool, such as Express, as this eliminates the need for multiple 4GL's;
4. The use of a domain-specific DSS generator, such as Excel, EFPM, or SAS packages. These are usually used where structured systems are to be built, most of the time in functional applications;
5. The use of CASE tools in development, such as Cool:Gen by Computer Associates. Suitable for strategic DSS;
6. To develop complex DSS use a combination of the previous approaches. This approach is suitable where the problem under investigation is complex.

The approach utilized for this research was to **develop the DSS using CASE methodology**. The reason is CASE supports the strategic management functions more than the others. The CASE methodology also ensures high quality systems and develops systems that are more responsive to user requirements (Laudon and Laudon, 2001; Cool:Gen manuals, 1997; Davids, 1992).

Devlin (1997: 27) said "It may be assumed that operational data consistency problems are solely historical, caused by immature approaches to application development. It then follows that widespread use of modern computer aided software engineering (CASE) tools will eliminate these problems in future applications."

Laudon and Laudon (2001: 345) said "CASE is the automation of step-by-step methodologies for software and systems development to reduce the amount of repetitive work the developer needs to do. Its adoption can free the developer for more creative problem-solving tasks."

1-12 Research limitations

1. The research findings apply only to the Admission and Registration IS that were running in the period of data collection (from June to December 2000);
2. The proposed ARDSS will be developed based upon the information needs of certain managers' positions to which this system is designed. The positions investigated were; Deans, Associate Deans, Registrars, Admission Officers, and Others whose positions enable them to take admission and registration-related decisions;
3. The ARDSSQ is an industry specific questionnaire. This means that it is only applicable to the higher education institutions;
4. The ARDSSQ is designed for the Egyptian Universities, however its use can be extended to other countries providing modifications are made to reflect the country-specific education system and regulations;

5. The ARDSSQ is only relevant for evaluating the Admission and Registration IS, not any other IS;
6. *Data limitation.* Some of the required decisions have not been implemented in the ARDSS because no data was available;
7. *Technical limitation.* As the development of software using CASE tools is a team work process, this has restricted the capabilities of the generated ARDSS. Although the ARDSS is working properly, it could have been enhanced if development teamwork was available.

1-13 Thesis plan

The thesis consists of eight chapters. A summary of each of the chapters follows:

Chapter one

Introduction

Chapter two

Overview of Decision Support Systems

This chapter will cover the information systems' details especially decision support systems (DSS). It starts by introducing the idea that information systems are built to resolve business problems. It then explores the idea that different information systems are used at different situations by different level of managers. The components the information system are discussed and the information system's competitive role is ascertained, within this context the strategic information systems (SIS) concept is explored. The different types of information systems are discussed including transaction processing systems (TPS), management information systems (MIS), expert systems (ES), office automation systems (OAS), artificial neural networks (ANN), and executive information systems (EIS), executive support systems (ESS), and decision support systems (DSS). The components of the DSS are introduced including; data management, model management,

knowledge management, user interface, and system user. The different classifications of DSS are also discussed. The DSS development processes are introduced including system development life cycle (SDLC), prototyping, and end-user computing. Also the group decision support system concept is highlighted, followed by the hybrid support systems. Finally, the chapter ends with comparisons between the different information systems.

Chapter three

Data warehousing

This chapter covers data warehouse (DW) details in the context of the DSS development. The chapter starts with introducing the different data sources in organisations; internal, external, archival and personal. Then the relationship between these data sources and the data warehouse is explored. The traditional database models; relational, hierarchical, network, object-oriented, and the multi-media are then illustrated because these are the models used to develop operational data stores (ODS). Then the definitions and features of the data warehouse are illustrated and a DW definition is proposed. The data warehouse characteristics are discussed including time-variant, non-volatile, subject-oriented and integrated. The benefits, both the tangible and intangible, of a DW are then elaborated. The differences between data warehouses, data marts, and enterprise data warehouses are discussed. The star schema structure is introduced in detail. The data warehouse components are then discussed followed by the Client/Server concept and its relationship with data warehousing. Top-down and bottom-up approaches to the development of a DW are introduced followed by defining the users of the data warehouse. Then, the DW development strategy is discussed. The chapter ends by suggesting some DW development guidelines.

Chapter four

Knowledge Discovery in Database Techniques

This chapter covers the knowledge discovery in database (KDD) process. The definition of the KDD process and its importance will be defined. Different terminology for the knowledge discovery process will be discussed with particular emphasis on data mining. The distinction between KDD and data mining will be clarified by showing the place of the data mining in the KDD process. The tasks, goals, and components of the data mining algorithms are illustrated. The data mining techniques including query tools, visualization, on line analytical processing (OLAP), association rules, decision trees and rules, artificial neural networks (ANN), clustering, genetic algorithms and probabilistic graphical dependency technique will be discussed. The role of the user in the KDD process will be discussed. To place the entire KDD process in context, it is applied to a sample data set drawn from the Arab Academy for Science & Technology and Maritime Transport (AASTMT) records. Finally, the chapter ends with illustrating some of the research and application challenges facing KDD.

Chapter five

The DSS, KDD, and DW blend of technologies

This chapter will establish the relationship between the DW, KDD, and DSS. It also introduces how these components will work together. The proposed DSS methodology is explored. Justification of the tools and mechanisms (i.e. the questionnaire, Cool: Gen CASE tools, and MS-SQL Server) employed to complete the methodology will be evaluated. Also, the data mining techniques used for the knowledge discovery process including: SQL, visualization, and cluster analysis will be analyzed.

Chapter six

Evaluating the current DSS in universities

In order to build the proposed DSS two steps had to be taken. Firstly is to evaluate the current admission and registration information systems in the Egyptian Universities. Secondly is to extract the users' requirements for a new system. A multi-part questionnaire was developed and distributed to these universities. This chapter starts by discussing the questionnaire development and then identifies the population and sample. After that it introduces the respondents distribution. Finally analysing the results and findings of the collected data according to the objectives of this dissertation.

Chapter seven

The proposed Admission and Registration DSS

This chapter covers the development details of the proposed Admission and Registration DSS (ARDSS). The chapter will start with the ARDSS development methodology. The proposed DSS methodology that was introduced in chapter five will be adopted in the development of the ARDSS. The methodology consists of four modules; module 0 to identify the users' needs, module 1 to build the data warehouse, in module 2 the KDD process is applied to 1800 sample data records, and in module 3 the DSS is being developed. The discovered knowledge will be discussed deeply in modules 1 and 2. The ARDSS will be implemented using Cool: Gen CASE tools and MS-SQL Server. The chapter will also discuss the management implications of the ARDSS . Finally, the relevant to the DSS development study objectives will be discussed within the chapter.

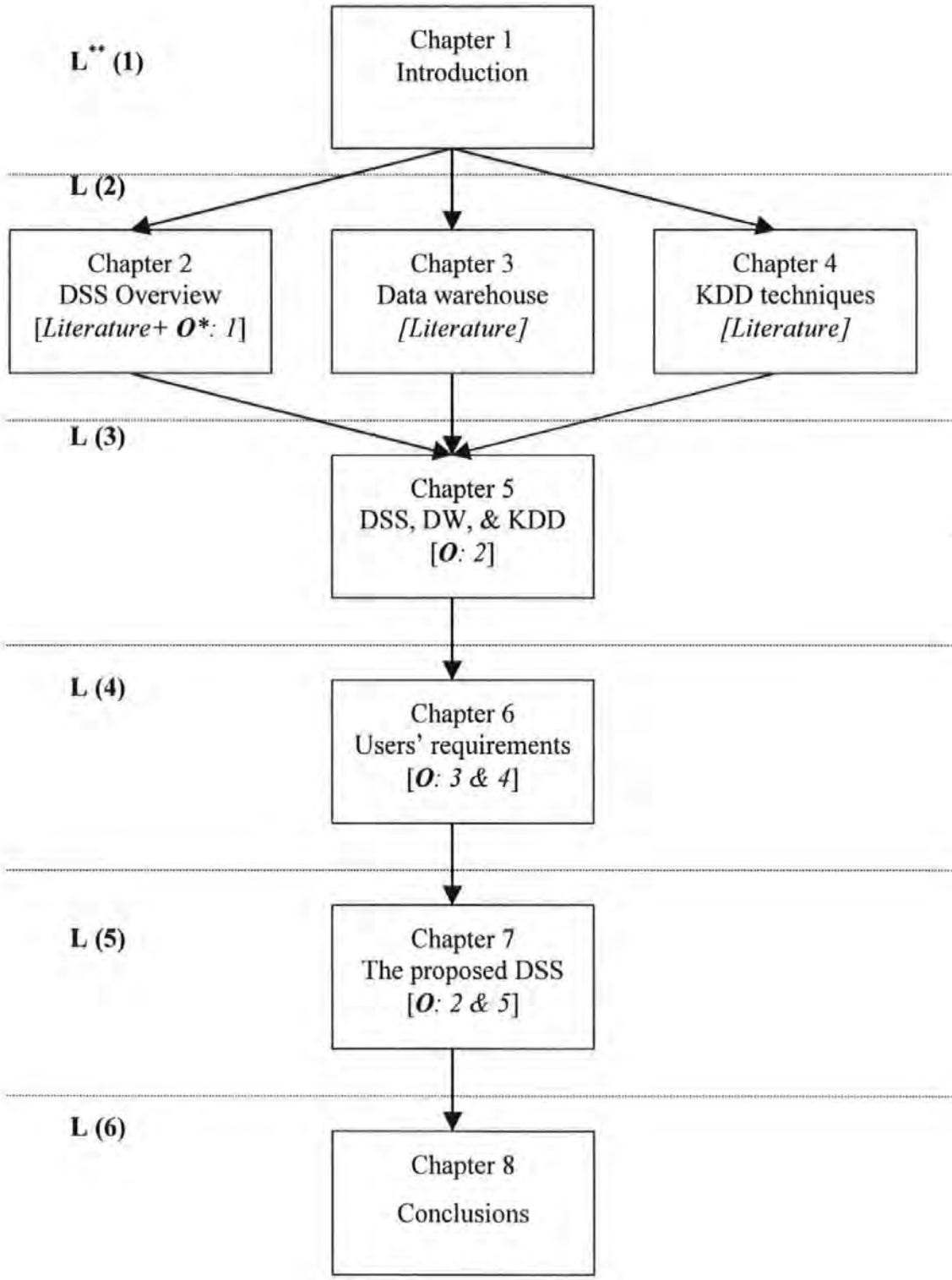
Chapter eight

Conclusions and Recommendations

Finally, it is worth mentioning that the thesis chapters are to be viewed as layers as follows:

- Layer (1) contains chapter one which is an introduction to the thesis and includes the study objectives;
- Layer (2) contains chapters two, three, and four which covers the literature review;
- Layer (3) contains chapter five which uses DSS, DW, and KDD to formulate a new DSS methodology and its mechanisms;
- Layer (4) contains chapter six which studies the population of this research in terms of the current systems users' requirements for a new system;
- Layer (5) contains chapter seven which achieves the users' requirements utilizing the new DSS methodology;
- Layer (6) contains chapter eight which is the thesis conclusion and recommendation.

The idea behind these layers is that each layer depends on the previous one and leads to the next. The following figure (1-2) describes this idea.



* O: Research objectives covered.
** L: Thesis layer.

Figure (1-2). A layered thesis plan.

Chapter two

DSS Overview

This chapter describes and evaluates the use of information systems especially decision support systems (DSS). It starts by defining the fundamental concepts of data, information, and knowledge. Then, investigates the different management levels in organisations and their needs and then explores the idea that different information systems are used in different situations at different levels of management. The components of an information system are discussed and the information system's competitive role is ascertained. Within this context the strategic information systems (SIS) concept is explored. The different types of information systems are discussed including transaction processing systems (TPS), management information systems (MIS), expert systems (ES), office automation systems (OAS), artificial neural networks (ANN), executive information systems (EIS), executive support systems (ESS), and decision support systems (DSS). The components of DSS are described including; data management, model management, knowledge management, user interface, and system user. The different classifications of DSS are also discussed. The DSS development processes including system development life cycle (SDLC), prototyping, and end-user computing are introduced. Also the group decision support system concept is highlighted, followed by the hybrid support systems. Finally, the chapter ends by comparing the different information systems based on different viewpoints.

2-1 Data, information, and knowledge

Data, information, and knowledge are three fundamental concepts which unfortunately are used interchangeably as synonyms whilst in fact they are not.

2-1-1 Data

Data are raw facts about things, events, activities, which are classified and recorded but not organised to convey a specific meaning. Data may be text, numbers, figures, sounds, or images.

Sometimes data are computer readable, e.g. numbers written on bank ckecks are being automatically read by input devices. However, most data does not exist in a form that can be read by the computer. For instance, lecturers record their students' grades on mark sheets and then users (i.e. registrar, secretary, or the lecturer himself) log to a certain program and start converting the data on those sheets to a computer format. In this case, those sheets are called *source documents* and the marks are called the *source data* (Long and Long, 2001).

2-1-2 Information

Information is data that has been processed and has specific meaning to someone. Someone's data provides the foundation of another person's information. For example the number of credit hours of a certain student is data for him whilst it is information for his lecturer.

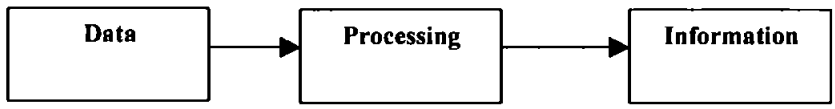


Figure (2-1). Data and Information.

Information is useful in making decisions because it is based on processed data and is the output of any data processing system (Hicks, 1993). Managers require information to take decisions (Corr, 1995; Hicks, 1993). In order to be used by managers, information should have these qualities (Long and Long, 2001; Corr, 1995):

1. *Clarity.* Clear information means easy to understand. The information recipient should be provided with information that is understandable to them;
2. *Timeliness.* It means that the information is being provided to the right person at the right time. Timeliness also refers to the time sensitivity of information. The same information could be of less or probably no value one day or one month later e.g. stock market and military information;

3. *Relevance*. Information should be provided to the relevant person, so if the manager is entitled to take production-related decisions then this information must be relevant to him, otherwise irrelevant;
4. *Correctness and completeness*. Correctness means that information should be free from errors, whilst completeness refers to the degree to which information is free from omissions;
5. *Frequency*. Information should be provided at an appropriate frequency to the decision maker. That is if the decision to be taken is related to a monthly function, the information should be also monthly provided.

Thus in terms of this research information is processed data that is characterized as being; clear, provided to the relevant person at the right time, error-free & omission-free, and submitted to the decision maker at the required frequency.

2-1-3 Knowledge

Knowledge is a combination of experience, accumulated learning, and information that have been organised and analyzed to be understandable and applicable to a specific decision situation (Laudon and Laudon, 2000). The collection of knowledge related to a specific decision situation is called a *knowledge base*. Each knowledge base has a specific *domain*. Examples of knowledge domains are dentistry, sales, production, and accounting (Turban and Aronson, 1998).

Figure (2-2) depicts the relationship between data, information, and knowledge with regard to their quantity. It shows that data is the largest in quantity, then information, and finally knowledge. While managers acquire data items from various sources (subordinates, peers, competitors, customers, authorities, Internet, and bulletins) they process part of these data items into information. Organising this information and mixing it with the managers'

experience and accumulated learning develops their knowledge base about certain decision situation.

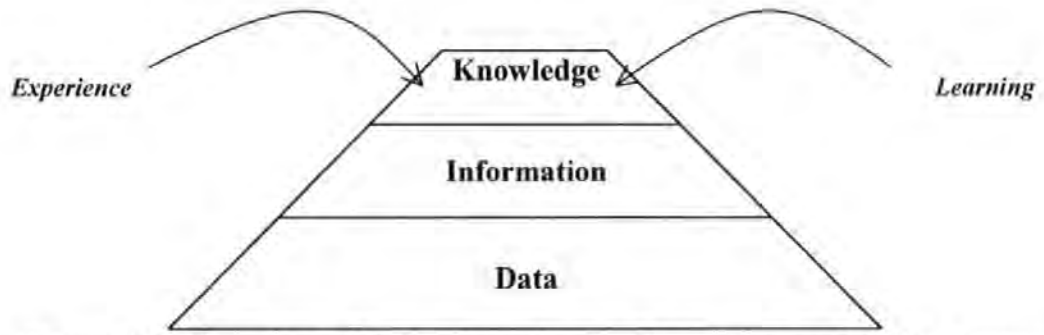


Figure (2-2). Quantities of data, information, and knowledge received by a manager[1].

The focus of this research is on how to deliver information and knowledge to business managers to better understand their business problems and hence to improve their decisions.

2-2 Management levels

The type and quantity of information required by managers vary with the level of management, e.g. operational, tactical, and strategic. Management at each of these levels is required to make different types of decisions and hence their information requirements are very different, these issues are developed in the following sections.

2-2-1 Operational level

Decisions at the operational level tend to be more structured. Structured decisions involve situations where the procedures to follow to reach a decision can be specified in advance. The inventory reorder decisions faced by most businesses are a typical example (Hicks, 1993). Managers at the operational level typically require regular internal reports detailing current and historical comparisons that support the structured control of day-to-day operations. Operational levels of management thus require pre-specified, frequently scheduled, and detailed information, with a more narrow, internal, and historical focus (O'Brien, 1996).

[1] Adapted from (Turban and Aronson, 1998).

2-2-2 Tactical level

Decisions at the tactical level are semi-structured. Semi-structured decisions are those that are partially well defined and the other part of the decision process is not well defined and is hard to predict. That is, some decision procedures can be pre-specified, but not well enough to lead to a definite recommended decision. At best, most decision situations are semi-structured. For example, decisions involved in starting a new line of products or making a major change to employee benefits would probably be classified as semi-structured (O'Brien, 1996).

2-2-3 Strategic level

Decisions at the strategic level are unstructured. Unstructured decisions involve decision situations where it is not possible to specify in advance most of the decision procedures to follow. The strategic management level requires more summarized, ad hoc, unscheduled reports, forecasts, and external intelligence to support its more unstructured planning and policy making responsibilities (Long and Long, 2001; O'Brien, 1996).

2-2-4 Comparison

Thus, we can generalize that higher levels of management require more ad hoc, unscheduled, infrequent summaries, with a wide, external, forward-looking scope. Table (2-1) summarizes the features of the three management levels.

Features	**Management Level**		
	Operational	**Tactical**	**Strategic**
Problem variety	Low	Moderate	High
Degree of structure	High	Moderate	Low
Degree of uncertainty	Low	Moderate	High
Time horizon	Days	Months	Years

Table (2-1). Features of the three management levels[2].

Thus, information systems must be designed to produce a variety of information products to meet the decision needs of managers at different levels in an organisation

2-3 What are Information Systems (IS)?

What makes IS different from the business systems is the information and information technology (IT) that IS use. Competition in between businesses is forcing managers to use information systems to cut their costs. Further, organisation that wish to remain competitive will seek employees who:

- Are educated in the use of technology;
- Can recognize potential applications for information systems technology and;
- Are capable of using IT in their day-to-day tasks.

Several definitions of IS have been proposed by Alter, Corr, Rowley, and Laudon & Laudon. These definitions will be stated together with a brief analysis.

2-3-1 Alter (1992)

Alter (1992) defined information system (IS) as being a combination of work practices, information, people, and information technologies organized to accomplish the organisation's goals. This is a very broad definition that encompasses things of different areas. The definition states that any IS consists of the following four components:

1. *Work practices.* These are the methods used to perform the daily work. They encompass not only procedures described in the manuals of operations but also approaches in which people coordinate, communicate and perform decision-making;
2. *Information.* Information is the output of data processing. And data are facts that are useful for a certain person at a certain time;
3. *People.* They are the providers of data and the users of information;
4. *Information technology.* All technologies that are employed by any IS; hardware, software, communications, and similar components.

[2] Adapted from (Hicks, 1993).

2-3-2 Corr (1995)

Corr states that IS refer to any computer-based systems which are used to assist in the management and operation of an organisation. The basic functions of IS are; collecting data, processing data, storing information and the dissemination of information to the decision maker. IS may be used in the following ways:

- Increase the efficiency (doing things right) and effectiveness (doing the right thing) of the business operations;
- Enable management to control the operations of the organisation;
- Improve the decision making process;
- Facilitate the co-ordination of activities within an organisation.

2-3-3 Rowley (1996)

Information systems are the collection, storage, processing, dissemination and use of information. IS are not only restricted to hardware and software, but are also related to the importance of man and goals of these technologies and the values employed in making these choices.

2-3-4 Laudon and Laudon (2000)

IS can be defined as a set of interrelated components working together to collect, retrieve, process, store, and disseminate information for the purpose of facilitating planning, control, coordination, and decision making in business and other organisations. These components are people, organisations, and technology.

They also said that each information system operates in a cycle of three steps; input, processing, and output, where there is a feedback from the output to the input step. In the input step the IS collects data from within the organisation and its environment. In the processing step the IS transforms the data into understandable and useful form. In the output step information is transferred to people that can use it. Feedback is the response of the people who

use the output of the IS which can be used to evaluate or modify the input step. Feedback can also be seen as data about the performance of the system. Figure (2-3) illustrates this process.

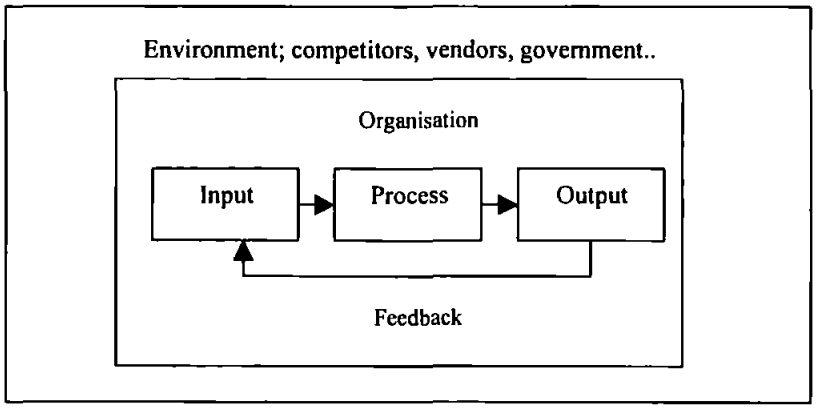


Figure (2-3). An information system[3].

2-3-5 Comparison

The above definitions of IS are not contradicting rather they are complementary. All of the definitions have the same components: hardware, software, data processing, information dissemination, work practices, and user managers but in different combinations and concentration. Laudon and Laudon have introduced a comprehensive IS definition that is not restricted to hardware and software but also encompasses human and organisation factors. An IS is not just about computer technology but how organisations and people employ the technology to facilitate their work and enhance their competitive position. For this reason their definition is the one that will be adopted in this research.

2-4 The Information Systems' competitive role

Information systems technology has increasingly becoming a vital part of any business strategy (Reynolds, 1995). So, since decision-makers and executives set strategies to gain competitive advantages in their markets they have to use information systems technology to enable them make competitive advances (Robson, 1997; Pegels, 1995; Keyes, 1993).

[3] Adapted from (Laudon and Laudon, 2000).

Providing value to customers or as sometimes-called *customer satisfaction* is the ultimate goal of any organisation (Czerniawska and Potter, 1998). Within any organisations products/services compete to constitute this value in different percentages. Managers inside the organisation try to use every weapon they have to increase the value of their products/services, and this what is called the *value chain* (Armstrong and Kotler, 2000). Organisations compete with one another based on their products' value chain. The product whose value is perceived to outperform those of its rivals will win the race (Alter, 1992).

An organisation can survive in the long-term if it develops successful strategies to confront the five competitive forces that draw the competition in the market of any industry (O'Brien, 1996; Geiger, 1992). These forces are *rivalry of competitors*, *threat of new entrants*, *threat of substitutes*, *the bargaining power of customers*, and *the bargaining power of suppliers*.

A variety of competitive strategies can be applied effectively to face these forces. An approach that is used by the organisations to face the five competitive forces is by employing five basic strategies (Pegels, 1995). The following outlines these strategies:

1. *Cost leadership*. Becoming a low-cost producer of service provider in the industry;
2. *Differentiation*. Introducing methods to differentiate the organisation's products and services from others';
3. *Innovation*. This involves introducing new products and services or entering new markets or attracting new customers;
4. *Growth*. Increasing the capacity of the organisation to produce products and services;
5. *Alliance*. Becoming linked to customers, suppliers, competitors, and consultants. These linkages might be merger, acquisition, joint ventures or any other marketing or distribution agreements.

Information systems are one of many approaches that an organisation can use to create a competitive advantage, and as a result of that making positive contribution to the product's value chain. However, managers must ensure linkages between the business and its IS projects. Managers also expect that when they are using an IS it has to be designed to make the decision making process more effective.

Utilizing information systems technology to improve the organisation's competitive position can be done in many ways. Organisations can use these methods to set their organisations apart from the competition, to sharpen their business strategies, maximize IS/IT (information technology) usage, improve their operations, and to reach competitive advantages (Keyes, 1993). Table (2-2) shows how to use information systems technology to implement the competitive strategies in order to face the competitive forces.

Competitive strategies	**Examples of how IS/IT can work**
Lower cost	-Use the IS/IT to reduce the cost of business. -Use IS/IT to lower the cost of customer or supplier related functions.
Differentiation	-Develop new features to differentiate the products or services.
Innovation	-Make radical changes to the business using processes. -Create new products and services that include the IS/IT components.
Growth	-Use IS to manage the global business expansions.
Alliances	-Develop inter-organisational IS to establish business relations with customers, suppliers, and others.

Table (2-2). How IS/IT can be used to implement competitive strategies[4].

The strategic role of IS involves the development of products and services and capabilities that give the organisation strategic advantages over the strategic forces in the market. The information system that performs this function is called a *strategic information system* (SIS),

[4] Adapted from (O'Brien, 1997).

or a *strategic management information systems* (SMIS) (Robson, 1997). SIS can be any kind of IS, it might be TPS, MIS, EIS, or DSS that helps an organisation achieve competitive advantage or reduce competitive disadvantage or meet the strategic objectives of the organisation (Robson, 1997; O'Brien, 1996).

2-5 Types of Information Systems

There are different types of information systems; transaction processing systems, management information systems, decision support systems, executive information systems, expert systems, office automation systems, and artificial neural networks (Long and Long, 2001; Laudon and Laudon, 2000; Reynolds, 1995; Turban, 1993; Alter, 1992). There is a need for each type of system and some managers will use an IS that would be not relevant or appropriate for other managers. Some of the determinants are the scope of decision effect, types of problems, future orientation, level of details, and many other factors that shape each IS.

2-5-1 Transaction Processing Systems (TPS)

In the 1960's the first IS were developed to automate manual transactions using electronic computers. These IS were called electronic data processing (EDP) systems but nowadays these systems are referred to, as transaction processing systems or fundamental information systems (Long and Long, 2001; Sprague and Carlson, 1982). Transaction Processing Systems record and collect data about the daily transactions taking place in any organisation. A transaction is a business event that generates or modifies data. The transaction is stored in an information system. TPS are also the backbone of any IS. Thus we can not build any IS, if we do not have a TPS (Reynolds, 1995; Alter, 1992). TPS are characterized by:

- Processing the detailed data necessary to update records about the fundamental business operations;

- The result records representing the organisation current situation;
- Data capturing system;
- Being highly structured but inflexible in design;
- An inability to give access to key information;
- The updating of data that may be captured interactively or in batches;
- Being based on a detailed model of how the transactions are processed;
- People who process transactions are the TPS users;
- Being short-term in nature;
- Having no built-in decision capability;
- Not providing management with information, they simply computerise the manual systems within an organisation.

2-5-2 Management Information Systems (MIS)

Management Information Systems provide the information required for managing the organisation. MIS have emerged in response to the shortcomings of TPS (i.e. the managers need for information). MIS are characterized by:

- Requiring TPS as a prerequisite;
- Summarizing data from one or more TPS and presenting it to the managers;
- Providing information but not how to evaluate performance, or how to take corrective actions;
- Targeting structured and semi-structured type of decision;
- Having a short-term orientation;
- Providing detailed output, not summarized, reports;
- Reading data from TPS;
- Focusing in providing information;

- Limiting access to authorized personnel only using system security procedures;
- Being primarily used by middle mangers;
- Providing scheduled and on demand reports.

2-5-3 Expert Systems (ES)

Expert Systems are designed to help managers make better decisions in certain areas. ES are interactive Computer Based Information Systems (CBIS) that respond to questions, make recommendations, and add value to the decision making process. ES are computerized advisory systems that try to mimic the reasoning process and knowledge of experts in a specific domain. ES are part of the evolutionary discipline of Artificial Intelligence (AI). The purpose of ES is not to replace the human experts, rather to make the rare expertise knowledge available to a wide range of people at the same domain (Turban and Aronson, 1998; Reynolds, 1995; Hicks, 1993). ES are characterized by:

- The specific domain of knowledge each individual system contains represents one of their advantages;
- Reducing the need for highly paid experts.
- Being able to be used as training tools for novices.
- Providing explanation as part of the decision making process;
- Focusing on the transfer of knowledge from experts to the system;
- Giving advice and explanation to their users;
- Being designed primarily for use by top managers and specialists;
- Having the ability to make complex decisions;
- Rules that are inferred from experts are kept in a knowledge base;

- The sources of ES problems are: the transfer of expertise from the expert to the ES, the definition of experts, and the contradictory rules given by different experts when building ES.

2-5-4 Office Automation Systems (OAS)

Office Automation Systems refer to all the CBIS associated with general office work or office activities. The activities performed by office staff in an organisation include; *managing documents*, *scheduling individuals and groups*, *managing data*, and *managing projects* (Laudon and Laudon, 2000). OAS allow the creation, storage, and communication of information in written, verbal, or video form throughout the organisation or between organisations (Corr, 1995). OAS applications are used to support the organisation activities including; word processing, e-mail, image processing, document copying, document image processing, voice processing, groupware, telecommunications, internet, desktop publishing, etc. Not all of these applications need to exist in one OAS. Office Automation Systems were thought to be primarily designed for secretaries and clerical workers, but this is only one aspect of the OAS. Looking at the applications that modern OAS support reveals that professionals, managers, clerical, and sales employees are all dependent upon an OAS to carry out their daily jobs (Laudon and Laudon, 2000). Office information systems are information systems that address traditional office tasks (Regan and O'Cnnor, 1994; Reynolds, 1995; Corr, 1995). OAS are characterized by:

- Facilitating every day communications of the office;
- Being oriented toward data rather than models;
- Being used by managers since their jobs include general office work;
- Primarily designed to satisfy certain function in the office;
- Helping people increase their personal productivity.

2-5-5 Artificial Neural Networks (ANN)

Martin et al. (1994) said that Artificial Neural Networks attempt to tease out meaningful patters from vast amounts of data. ANN can recognize patterns, and they can adapt as new information is received. The key issue to ANN is that they have the ability to learn. ANN are given a database consisting of many variables concerning certain circumstances. They analyze this data to find correlations between the variables. Using this structure the ANN attempts to predict the outcome in certain cases. Examples of ANN are the area of pattern recognition and medical diagnosis. ANN are characterized by:

- The ability to deal with incomplete information;
- Generating new knowledge as new information is received;
- Being applications of AI.

2-5-6 Executive Information Systems (EIS)

Alter (1992) said that Executive Information Systems are highly interactive systems that provide managers and executives with flexible access to information for monitoring operating results and general business conditions. A key issue in the definition of an EIS is that it is designed for the executives to use without any aid from intermediaries. An EIS uses state-of-the-art graphics, communications, and data storage methods to provide executive with access to the current situation of what is happening in the organisation (Martin, et al., 1994). An EIS provides the executive with an electronic window to look at what is happening in the organisation. The benefits of using EIS are many:

- They provide the executives with the key information in a short time;
- They eliminate communication obstacles between the executive and his peers, or subordinates (Houdeshel and Watson, 1992; Corr, 1995);
- The executive is expected to be user-influential, rather than user-responsive (Ball, 1992);

- EIS provide competitive advantages (Reynolds, 1995).

2-5-6-1 Elements of successful EIS

To support the pre-defined roles of the executive EIS should encompass the following elements (Reynolds, 1995).

1. *Standards reports.* EIS should have the capability to navigate easily through large amounts of data to create standard reports;
2. *Short term issues.* Most of the time executives are faced with short-term problems. Areas where short-term problems might be found are inventory control, customer billing, order processing and production scheduling. The role of the EIS is to handle these short-term issues through spreadsheets, graphs, analysis, and to prioritise them;
3. *Exception reporting.* This feature enables the executive to determine the criteria of the exception reports via a menu of identified exceptions;
4. *Executive brief.* Executives should have the ability to download what is of interest to them to their local machines, and tailor the system to meet their common tasks;
5. *External data.* The EIS should be able to handle external data to enrich the analysis and decision making associated with it;
6. *News.* There has to be an automatic delivery of news to the executives. Some of this news is in the field where the executive works whilst other news is of a more general nature;
7. *Data analysis.* An EIS should include different tools and mechanisms of analysis from simple calculations to the complex models;
8. *Executive mail facility.* This is a very useful facility through which the executive can send or receive a screen of data and wait for a response or reply;
9. *Time management.* This is an electronic calendar to keep the manager in touch with the meetings, visits, tasks, etc;

10. *Information retrieval.* Executives should be able to access aggregated information about the organisation. For example the executive retrieves the total sales volume without being able to identify how regions contribute to this number, or what is the percent of each product by sales person to this number;

11. *Drill down capability.* The reports' capability of the EIS should offer the facility to the executive to drill down to detailed data. That is, to provide the details of any given information. To provide such a capability the EIS may include several menus, and submenus (Turban, 1993). For example the executive is able analyze the volume of sales by identifying each product's contribution, sales persons, and regions to the total sales ether in units or in money terms;

12. *Internet connection.* This is to reflect the trend in business; for example making deals through the Internet (Laudon and Laudon, 2000; Reynolds, 1995).

2-5-6-2 Executive Support Systems (ESS)

Whilst a great number of researchers use the two terms-EIS and ESS- as synonymous (Laudon and Laudon, 2000; Corr, 1995), others distinguish between them (Rockart and Delong, 1988) with Executive Support Systems having a wider definition (Rockart and Delong, 1988). In their definition, ESS includes the following capabilities- in addition to those which an EIS offers:

- Support of electronic communications. Example: E-mail;
- Data analysis capabilities. Example: Spreadsheets;
- Organizing tools. Example: Calendar.

However, for the purpose of this research EIS and ESS are synonymous as in real life applications EIS provide the ESS capabilities. For example many of the EIS development efforts include E-mail and Calendar.

Kumar (2000) introduced another type of executive information systems called Global Executive Information Systems (GEIS). According to Kumar, these GEIS are defined as being CBIS, provide access to internal and external data, used to support senior executives with analysis and decision making functions, and are only used by senior executives at headquarters in a global organisation.

2-6 Decision Support Systems (DSS)

2-6-1 Definitions

As is the case in most of the definitions in the field of information systems, there is no consensus on what DSS are, rather there are contributions from many researchers. In order to reach an acceptable working definition these definitions will be introduced and then analysed.

2-6-1-1 Scott-Morton (1970)

The emergence of DSS and related issues occurred in the early 1970s by Scott-Morton under the title management decision systems (Turban, 1993). Their definition states that DSS are interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems.

2-6-1-2 Little (1970)

Little (1970) defines DSS as model-based set of procedures for processing data and judgments to assist a manager in his decision making. Little also said that for the DSS to be successful it must be simple, robust, controllable, adaptive, complete, and easy to communicate with.

2-6-1-3 Alter (1980)

Alter (1980) defines DSS by contrasting it with traditional electronic data processing (EDP) systems as in the following table (2-3).

Dimensions	DSS	EDP
Use	Active	Passive
User	Line and staff managers	Clerical
Goal	Effectiveness	Mechanical efficiency
Time horizon	Present and future	Past
Objective	Flexibility	Consistency

Table (2-3). DSS against EDP[5].

2-6-1-4 Bonczek et al. (1980)

Bonczek et al. (1980) defined DSS as CBIS consisting of the following: a language system, knowledge system, and a problem processing system.

2-6-1-5 Keen (1980)

Keen (1980) defines DSS as a situation where a final system can be developed through an adaptive process of learning and evolution.

2-6-1-6 Sprague and Carlson (1982)

Sprague and Carlson (1982) defined DSS as being dedicated to improving the performance of knowledge workers in organisations through the application of information technology. The definition focuses on four issues; improving performance, knowledge workers-the decision makers in organisations, organisations, and the application of information technology.

2-6-1-7 Bennett (1983)

Bennett (1983) defined DSS as coherent systems of computer based technology used by managers as an aid to their decision making in semi-structured and unstructured tasks to support rather than replace managerial judgments, focusing on improving the effectiveness rather than improving their efficiency. In this definition efficiency means doing things right the first time, however, effectiveness in the decision making process addresses the problem of what should be done or doing the right thing. Bennett also said that managers should be concerned with efficiency if they are to resolve structured problems and to be concerned with

effectiveness if they are to resolve semi-structured or unstructured problems. DSS can be used to address semi-structured and unstructured problems. Hence the use of a DSS should enhance the manager's effectiveness.

2-6-1-8 Stevens (1991)

Stevens (1991) said that the appearance of DSS was due to the increased needs of the organisations for information that the MIS failed to provide. For example handling the complex problems and the need to give some recommendations and evaluate decision options. He also said that the emergence of the DSS was in subsequent stages:

1. *Gain accesses to corporate data.* Here the DSS can not deal with the organisation transactions, so DSS are incapable of reaching decisions based on these data. This problem has been resolved by making a link between the organisational data and the DSS by establishing *datapools*. The data pools represent the corporate databases;
2. *What-if models.* The ability to analyze the organisational data using spreadsheets and some financial modelling software has lead to the development of what-if and scenario models. In these models the manager can ask some what-if questions about a situation and DSS respond with the effect should this situation happened. Examples include the effect of cutting a product price on the total sales, the effect of a 10% salary increase on budget, forecasting optimum production mix as production units change, etc. These models are deterministic, that is they are constructed based on some input rules.
3. *More sophisticated models.* These models study the characteristics of data and allow inferences to be made. These models are either statistical or probabilistic (they are not deterministic). Examples of where these models may be applied include segmentation, profitability, sales forecasting, and performance modelling.

[5] Adapted from (Turban, 1993).

4. *Knowledge-based models*. Some complex decisions require combining the knowledge of experts with mathematical models to be resolved. Examples where these models arise are the pricing problem and the resource allocation problem.

2-6-1-9 Corr (1995)

Corr (1995) defined DSS as interactive systems which use data and models to help in the decision making process. DSS are designed for senior management although managers at all levels may find them useful. The main objective of DSS is to improve the effectiveness of decisions.

2-6-1-10 Reynolds (1995)

DSS are CBIS used to help people reach decisions. DSS can be applied to support operational, tactical, or strategic decision making. DSS can provide access to both corporate data and external data related to the problem being studies. The data can be used as input to a model to simulate the real world and display the result in different ways including graphs.

2-6-1-11 O'Brien (1996)

O'Brien (1996) defines DSS as information systems that use analytical models, specialized databases, decision maker's judgment in an interactive way to support the process of taking semi-structured and unstructured decisions by individual managers.

2-6-1-12 Marakas (1998)

Marakas (1998) defines DSS as information systems that can be identified by having three basic features; the problem structure (i.e. unstructured problems), the decision outcomes (DSS produce quality decisions), and the managerial control (can be used by top managers).

2-6-1-13 Long and Long (2001)

Long and Long (2001) defined DSS as interactive information systems that rely on an integrated set of user-friendly decision support tools (hardware and software) to produce information to support management in the decision making process.

2-6-1-14 Discussion of the study's first objective No. 1 "Investigate and critically evaluate the current DSS practices"

It is true that past studies on DSS have made significant contributions, however, new contributions must address their shortcomings due to the following developments. Technological advances have put new demands on IS and make even more sophisticated demands on computer support. Businesses are facing tough competitive pressures and organisations are becoming increasingly dependent on the successful use of computerised information systems.

The past DSS research has focused on many issues, but no integrated approach has been found because each study has tried to narrow the different aspects of the DSS. This focus is summarized in table (2-4).

Researcher	Definition focus
SCOTT-MORTON **70**	CBIS Data utilization Unstructured problems
LITTLE **70**	Data processing Model-based Judgement
ALTER **80**	Effectiveness Flexible
BONCZEK **80**	Language system Knowledge-based system
KEEN **80**	Adaptive system
SPRAGUE AND CARLSON **82**	Knowledge workers Performance improvement
BENETT **83**	CBIS Effectiveness Semi and unstructured problems
STEVENS **91**	Corporate data Models Knowledge-based system
CORR **95**	Data and models Senior management Effectiveness
REYNOLDS **95**	CBIS All management levels Internal and external data
O'BRIEN **96**	Analytical modelling DB Semi and unstructured Judgement
MARAKAS **98**	Problem structureness Decision outcomes Managerial control
LONG AND LONG **01**	Interactive information systems User-friendly Supporting the decision making process

Table (2-4). DSS definitions' focus.

Studying these summaries reveals that each of the DSS definitions proposed have some general characteristics. For instance, during the 70's the DSS definitions (Scott-Morton; Little) focused on data processing, and model. During the 80's the definitions (Alter; Bonczek; Keen; Sprague and Carlson; Benett) focused on CBIS, effectiveness and the knowledge component appeared as part of the DSS. Whilst during the 90's the DSS definitions (Stevens; Corr; Reynolds; O'Brien, Marakas) focused on the models and problem structure with more attention given to the knowledge component. Recent definitions (Long and Long) required the DSS to be interactive and user-friendly. However, no definition has been found to include all the following aspects: the type of data used, the management level, the DSS effect, effectiveness of the DSS, type of knowledge targeted by the DSS.

The proposed DSS definition that will be adopted by this research will be introduced in chapter five, because the definition components include items that will be elaborated in chapters three and four.

2-6-2 Why use DSS?

A study done by Houge and Watson in 1983 defined six major reasons why corporations employ large-scale DSS. This is summarized in table (2-5) below.

Reasons	Cited by %
Accurate information is needed	67
DSS are viewed as an organisational winner	44
New information is needed	33
Management mandated the DSS	22
Timely information is provided	17
Cost reduction is achieved	6

Table (2-5). Reasons for using a DSS

Research by Udo and Guimaraes (1994) that used 201 U.S. companies showed that the benefits of using DSS are higher decision quality, improved communication, cost education,

increased productivity, time savings, and improved customer and employee satisfaction. Also the degree of competition, the industry, size of the company and user-friendliness of the DSS were found to be highly correlated with the perceived benefits of DSS.

2-6-3 DSS characteristics

DSS can address very complex real problems. The major distinction between DSS, and traditional TPS and MIS is that DSS have the ability to use simulation models under the control of the user manager. (Reynolds, 1995; Turban, 1993; Alter, 1992). DSS are characterized by:

- They can be applied to support operational, tactical, and strategic level problems;
- Assisting managers to make repetitive decisions;
- Helping managers to evaluate options and choose the best one;
- Working within a short-term frame;
- Handling complex problems where a lot of data needs to be analyzed to support the decision making process. For example Airline DSS that analyse data collected on aircraft utilization, seating capacity and utilization, aircraft statistics, forecast airline market share, aircraft assignment, route requests, ticket classifications, and revenue and profitability;
- The need for an interaction between the decision-maker and the DSS;
- Being able to reach a recommendation sooner;
- Being developed by non-data processing (DP) professionals;
- Focusing on the flexibility of decision making;
- Providing information to support certain decision area;
- Being designed to support decision-makers at all levels, but they are most effective at the tactical and strategic levels.

2-6-4 Components of DSS

A DSS is composed of five parts (Marakas, 1998; Turban and Aronson, 1998). These are:

1. Data management subsystems;
2. Model management subsystems;
3. Knowledge management subsystems;
4. User interface subsystems;
5. The user.

2-6-4-1 The data management subsystem

The data management subsystem consists of four components; the DSS database, the database management system (DBMS), the data directory, and the query facility.

1. *The Database (DB)*. A DB is a collection of related data that have a common purpose (Elmasri and Navathe, 2000). DSS may have one or more databases (Turban and Aronson, 1998).

The databases included in the DSS might be internal, external, archival or private. An internal DB belongs to the organisation and contains its transactions, like payroll, sales, and inventory DB. An external DB contains data sources external to the organisation. For example data about competitors, customers, or vendors. An archival DB stores the organisation historical data. The private DB belongs to one or more of the DSS user managers.

When DSS are used to investigate ad hoc problems data can be entered directly to the DSS without the need for an independent DB. At other times, data can be obtained directly from an existing DB. To create a DSS database or data warehouse, it is often important to capture data from many sources. This process is called data extraction. Another extraction activity happens when users produce reports from the DSS database. The extraction process is totally dependent on the DBMS.

Some DSS databases are centrally stored in one basket called the data warehouse (DW). The DW includes the organisation's historical data across years. The combination of DSS and DW gives managers a strategic tool that will enhance the decision making process. Further analysing the organisation historical data increases the possibility of finding unknown facts and hidden information and patterns. Thus the use of a DW will enhance the strategic use and value of the DSS.

Both EIS and DSS can work as a front-end tool that is able to utilize the output of the DW in an efficient manner (Berson and Smith, 1997; Adriaans and Zantinge, 1996; Taha, et al., 1997; Paller, 1997; Barquin, 1997; Inmon and Hackathorn, 1994).

2. *The DBMS.* The DBMS is a software package that is used to create and maintain a database (Elmasri and Navathe, 2000). The following summarizes the capabilities of the DBMS within a DSS:

- Extracts data from the DSS DB;
- Includes the record maintenance functions of add, delete, read, update, search and print. However, if the data source is a data warehouse, not an operational database, only the search and print processes are applied;
- Obtains and processing interrelated data from different sources;
- Retrieves data from the DB;
- Provides a secured data source;
- Manages the data through data directory.

3. *The data directory.* The data directory or as sometimes called data dictionary or catalog includes data about data. It includes the data definitions and its functions. The directory supports the data maintenance functions. For example add, edit, print, delete, and retrieve information on certain objects.

4. *The query facility.* One of the most important functions in the DBMS is to access and manipulate data. These functions are accomplished by the query facility. The query facility using its query language accepts requests from the other DSS components and returns the required results.

2-6-4-2 Model base management subsystems (MBMS)

The model base management system is the software package that includes financial, statistical, operational, management science and other quantitative models. Its main purpose is to provide the DSS with its analytical capabilities. The MBMS consists of four components; the model base, the modelling language, the model directory, and the model execution and integration command processor.

1. *Model base.* The model base provides the analysis capability of the DSS. It is important to realize that the DSS can contain one model in some situations and up to several hundreds in others. The model base contains routines and models of special types through which the problem can be resolved and analyzed (Turban and Aronson, 1998). The model base can also include some model building blocks, which can be used as components in other models or can work independently. Examples of models include net present value calculations, and earning per share.

Models may also be classified by functional areas as well as by the management level. The models in the model base can be classified into four types: strategic, tactical, operational, and model-building blocks and routines.

The Strategic models are used to support top management functions like planning. Strategic models tend to be broader than the other models in spectrum and embody many variables and often use external as well as internal data.

Tactical models are used by middle level managers to assist them in the resource allocation and control related problems. Some external data may be required but the main data source is internal.

Operational models are utilized to support the analysis of routine and frequent problems. These models normally use internal data sources. The first-line managers are the fundamental users of these models. Examples include the processing of the daily work order.

The functions of MBMS are model creation, updating and model data manipulation (Turban and Aronson, 1998). Other functions of MBMS are to:

- Create models from scratch or from existing models;
- Conduct sensitivity analysis from what-if analysis to goal seeking;
- Maintain models;
- Catalog and display the directory of models;
- Use multiple models to support the analysis;
- Relate the models with the DB within the DSS.

2. *Modelling language.* A DSS deals with both semi-structured and unstructured problems, and with problems that are generally different from one user to another and from one organisation to another. Thus it is very important to have tools that enable users to customize their models. Using general purpose languages e.g. COBOL, BASIC, or special modelling languages e.g. IFPS (interactive financial planning system).
3. *Model directory.* The function of the model directory is similar to that of the data directory. The model directory contains the model description, its main functions, and its capabilities.
4. *Model execution, integration, and command processor.* Model execution is used to control the model whilst it is in use. Model integration is the process of integrating one or

more models in one problem. And a model command processor is utilized to accept and interpret instructions and send them to the MBMS (Turban and Aronson, 1998).

2-6-4-3 The knowledge management subsystem

This is the component that manages the knowledge or experience in the DSS. Experience is an optional component that can be added to the DSS to facilitate unstructured or semi-structured decision situations. Experience is provided to the DSS through an expert system (ES). Advanced DSS that require an ES are called intelligent DSS, or expert support systems, or knowledge-based DSS. The existence of the knowledge management subsystem is optional because some DSS do not require this feature. The current generations of the data mining applications all include such a subsystem (Marakas, 1998).

2-6-4-4 The user interface

The user interface is the medium between the user of the DSS and the DSS itself. This not only includes hardware and software configurations but also the ease of use and the manager's preferences. Whitten and Bently (1997) stated that the interface is a vital component because this is the only component that the user sees. Lack of ease of use or an inappropriate interface will result in users not using the systems no matter the depth and quality of the analysis procedures they offer. The user interface is managed by the user interface management subsystem (UIMS) which:

- Provides the graphical user interface (GUI);
- Handles a variety of input devices;
- Presents different format of data and for different devices;
- Gives help;
- Provides interaction with the database and the model base;
- Stores input and output data;
- Provides training by example;

- Interacts with dialog styles;
- Produces output reports.

2-6-4-5 The user

This is the decision maker who uses the DSS as a tool to enhance his information about the situation(s). A DSS has two classes of users: managers and staff specialists. Staff specialists include financial analysts, production planners, and marketing researchers. Managers tend to expect more user-friendly applications than staff specialists. Staff specialists tend to use more complex systems than managers. Users of the DSS must affect the systems development, for example their experience, education, style, preferences and area of the problems they face (Marakas, 1998; Turban and Aronson, 1998).

2-6-5 DSS hardware requirements

DSS have evolved concurrently with the advances in hardware. Hardware affects the functionality and the capabilities of the DSS, however, it also happens that the hardware is determined by what is available inside the organisation. In these situations it is very important to consider the balance between what we expect from the DSS and what the existing hardware configuration can support (Thierauf, 1988).

2-6-6 Classifications of DSS

DSS can be classified according to many factors, for example the extent to which the system outputs directly support the decision, the DSS orientation, type of problems, the degree of non-procedurality, and the type of DSS support. These classifications are discussed in the following sections.

2-6-6-1 Donovan and Madnick (1977)

1. *Institutional DSS.* This type of DSS deal with the routine and frequent problems in organisations. Examples include evaluating investment opportunities, which may be built incrementally across years;
2. *Ad hoc DSS.* Deal with specific problems that are not recurring or frequent. For example the planning and budgeting decisions.

2-6-6-2 The Taxonomy of DSS by Alter (1980)

This classification is based on the degree of action implication of system outputs, or in other words the extent to which the system outputs could directly support the decision. DSS are classified by their *orientation* and *class*. There are seven *classes* of DSS. The first two are *data oriented*, the third handles both *data and models*, the remainder are *models oriented.* This is illustrated in table (2-6).

Orientation	Classes	Type of Operation	Type of Task	User	Usage Pattern	Time Frame
Data	-File drawer	-Access	-Operational	-Non-managerial line personnel	-Simple	-Irregular
	-Data analysis	-Ad hoc analysis	-Operational, analysis	-Staff analyst or managerial line	-Manipulate & display	-Irregular or periodic
Data or Models	-Analysis information systems	Ad hoc Analysis, Multiple DBs	Analysis, Planning	Staff analyst	Special reports, Small models	-Irregular on request
Models	-Accounting	-Standard	-Planning, budgeting	-Staff analyst, or manager	-receive estimates	-Periodic
	-Represen-tational	-Estimates	- Planning, budgeting	- Staff analyst	-receive estimates	- Periodic
	-Optimiz-ation	-Calculation	-Planning, resources allocation	- Staff analyst	-receive answer	- Periodic
	-Suggestion	-Suggested Calculations	-Operational	-Nonmanagerial line personnel	-receive suggestion	-Daily, Periodic

Table (2-6). The Output-based classification of DSS[6]

2-6-6-3 Bonczeck et al. (1980)

The degree of non-procedurality is the main determinant to classify the DSS, so most builders have found that non-procedural languages (4GL) are faster and much convenient for DSS developments, unlike the procedural languages such as COBOL or BASIC.

2-6-6-4 Hackathorn and Keen (1981)

1. *Personal support.* The support here is given to an individual taking a decision;
2. *Group support.* The support is given to a group of people who are engaged in separate decisions but the decisions are correlated;

3. *Organisational support.* The focus here is the organisational tasks in a sequence of operations or different location and resources.

2-6-6-5 Holsapple and Whinston's Classification (1996)

1. *Text-oriented DSS.* Decision makers should be able to access the corporate stored databases which are always in a textual format. As data is accumulated so the amount of text to be searched by the decision maker increases. Thus, there is a need for text to be stored and processed efficiently. A text-oriented DSS helps the decision maker by allowing the document to be electronically created, revised, indexed, and processed as needed. Technologies that might be utilized to build a text-oriented DSS are document imaging, hypertext, and intelligent agents;
2. *Database-Oriented DSS.* In this type of DSS, the concentration is on the structure of the DSS itself. Normally database-oriented DSS are based on a relational structure, however, other database structures can be used such as object oriented or multi-media database structures. The main feature of this DSS type is the report generation and querying facilities where their power originates from the database technology;
3. *Spreadsheet-oriented DSS.* A spreadsheet is a modelling language that allows the user to write models to execute the DSS analysis. Spreadsheets are widely used by end-users, the most common examples are Microsoft Excel, Lotus 123, and Q-Pro. Spreadsheets also interface with some Database Management Systems (DBMS) which allow the access to some of the database power functions such as reporting;
4. *Solver-oriented DSS.* A solver is a technique or computer program written to resolve a certain computation or a particular problem. Examples are the reorder level in a stock control system, optimum process settings, etc. A solver can be commercially pre-

[6] Adapted from (Turban and Aronson, 1998).

programmed in development software and examples can be found in Microsoft Excel. The solver-oriented DSS can be flexible by allowing the techniques to be modified or deleted;

5. *Rule-oriented DSS.* The knowledge component of the DSS includes rules, these rules are either qualitative or quantitative. Rules are the principal components of the knowledge base DSS; it extends the capabilities of the computer far beyond the data or model-base;
6. *Compound DSS.* It is a hybrid system that includes two or more of the previously stated DSS types. A compound DSS could be built by grouping a set of individual DSS each in one area of the decision situation;

2-6-6-6 Summary

Different DSS classifications have been introduced by different researchers, the following table (2-7) summarizes these classifications.

Researcher	DSS Classification
Donovan and Madnick 1977	Institutional Ad hoc
Alter 1980	Data-oriented Models-oriented Data and models
Bonczeck et al. 1980	Non-procedural languages-based Procedural languages-based
Hackathorn and Keen 1981	Personal support Group support Organisational support
Holsapple and Whinston 1996	Text-oriented Database-oriented Spreadsheet-oriented Solver-oriented Role-oriented Compound

Table (2-7). Summary of DSS Classifications.

2-6-7 The DSS development process

There are different methodologies through which a DSS can be developed, including the systems development life cycle, prototyping, and end-user computing. These are discussed in the following sections.

2-6-7-1 The System Development Life Cycle (SDLC) approach

The SDLC process may be considered to contain eight development steps. However, not all DSS go through all of these steps. (Meador et al., 1998; Keen and Scott Morton, 1978).

1. *Planning.* This step deals with the needs to the systems, and the problems that the DSS will be designed to address. It also investigates the key decisions that need to be supported by the DSS. This step should also handle the system feasibility in terms of financial and technical constraints (Whitten et al., 1994);
2. *Research.* This investigates the user requirements and compares them against the available resources. The DSS environment is checked during this activity;
3. *System analysis and conceptual design.* Determines the best development approach in developing the DSS and identifies the resources required to be used;
4. *Design.* The overall structure of the systems components, architecture, and DSS features;
5. *Construction.* The construction depends on the design of the DSS. Once the system is constructed it should be tested and approved;
6. *Implementation and user training.* The DSS is evaluated to see how close it is to what was required and demonstrating the full functionality of the system. This activity shows the user how to use the system, training the users in real life session, and finally the deployment of the DSS to the users;
7. *Maintenance.* Continuous support and ongoing help provided by the DSS developers to the users. Proper documentation is also required and maintained as long as the system is in use;

8. *Adaptation.* This is the step of responding to the user requirements in the future whilst the system is running.

2-6-7-2 Prototyping

Since most DSS address semi-structured or unstructured problems most of DSS development is done by prototyping. Prototyping builds the DSS through a set of steps with spontaneous feedback from the user manager to ensure that the development process is running on the proper track. Prototyping is sometimes called the evolutionary approach or the iterative process or just prototyping (Turban and Aronson, 1998). The iterative process includes the following four tasks:

- Select an appropriate subproblem to be built first;
- Develop a small but usable system for the decision maker;
- Evaluate the system constantly;
- Refine, expand, and modify the system in cycles. Here the analysis, design, implementation and evaluation phases are repeated until the required system is reached.

The advantages of using prototyping are:

- Short development time;
- Short user reaction time;
- Low cost;
- And the improved user understanding of the DSS (Reynolds, 1995).

On the other hand, the disadvantages of the prototyping come from the loss of the advantages of utilizing the SDLC approach; detailed description, thorough understanding of the system, well tested system, and easily maintained. However, the risks of the prototyping can be reduced if merged with the critical success factors technique (CSF). The CSF is any process or procedure if performed successfully will ensure the success of the organisation as a whole (Rockart and DeLong,1988; Volonino and Watson, 1992).

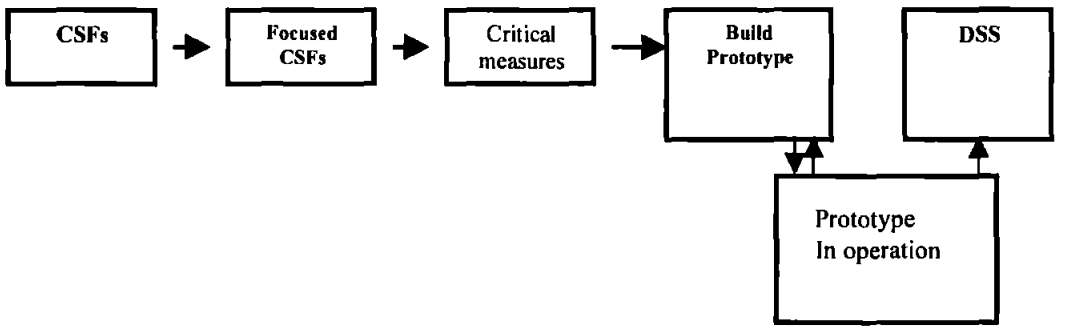


Figure (2-4). Combining Prototyping with CSF's[7].

2-6-7-3 End-user computing

End-user computing, which is also known as end-user development, is the development of CBIS by people outside the formal information systems area e.g. managers, professional users (financial analyst, engineer, and lawyer). They build DSS to support their work and enhance their productivity (Turban and Aronson, 1998). The end-users can be at any organisational level. The number of end-users has increased dramatically during the past decade due to the rapid growth of distributed computing and the use of Internet. A study by Mittan and Moore (1984) indicated that some of the top managers like to build their DSS themselves.

-*Advantages* for end-user development include short delivery time, eliminating the user requirement obstacles including the lack of communication or understanding. Further, there is no need for user training because the user is the developer of the system.

-*Disadvantages* for end-user development include poor quality systems, lack of user experience, inability to develop workable systems, loss of data, and lack of documentation and maintenance (Whitten, et al., 1994).

[7] Adapted from (Turban and Aronson, 1998).

2-7 Group Decision Support Systems (GDSS)

Within an organisation it is either an individual or a group that makes decisions. When decisions are to be taken by a group GDSS are used. GDSS allow a variety of specialists to be assembled whereby each of them is contributing to the solution using his expertise. GDSS could stimulate creative thinking and allow people from different departments to take the decision together. The disadvantage of GDSS is the possible conflict of the people contributing to the solution because each has a departmental view, but when all of them share the organisational goals this could encourage them to be committed to the organisation rather than individual departments (Lee, at al. ,1999).

Many attempts have been made to improve the work of a group, for example groupware, electronic meetings, collaborative systems, and GDSS. These systems are also known as group support systems (GSS) (Turban and Aronson, 1998; O'Brien, 1997).

2-8 Hybrid Support Systems

All computer-based information systems (CBIS) share the same objective that is to assist managers in their decision making, or in other words to transfer the managers from an *uncertain* situation into a situation of *certainty* or *risk* (any point between certainty and uncertainty) by providing complete or partially complete information. To complete this process one or more information systems might be used. A study conducted by Forgionne and Kohli (1995) found significant enhancements when integrated systems are used. The key point is to resolve the managerial problem not the use of a specific tool or system (Turban and Aronson, 1998).

When information systems are used together, many approaches can be employed. For example:

- Employ each tool to resolve part of the problem;
- Employ several tools that are loosely integrated;
- Employ many tools in an integrated manner.

Taking into considerations that these tools might belong to different vendors it is essential that these systems should be compatible. For example the ability to use common file formats e.g. ODBC (object database connectivity) and/or to use standard components e.g. SQL (structured query language).

2-9 Comparisons between different ISs

Table (2-8) summarizes and compares the different information systems discussed.

IS	What the system does	Focus	Users
TPS	Collect and stores information about transactions.	Data transactions	People who process transactions.
MIS	Convert data from TPS into meaningful information for managing organisation.	Information	Managers who receive feedback about their work.
DSS	Help people make decisions by providing information, models, or tools for analysis.	Decisions	Analysts, programmers, and professionals.
EIS	Provide information in readily accessible, interactive format.	Tracking and control	Executives and high level managers.
ES	Make the knowledge of experts available to others.	Transfer of expertise	People who solve problems.
OAS	Provide tools to make general office work more efficient and effective.	Increase worker productivity	Office workers.
ANN	Learn from incomplete and partial information.	Pattern recognition	Managers, and professionals.

Table (2-8). Information Systems overview[8].

Laudon and Laudon (2000) have examined the differences between IS in the light of Henry Mintzberg research. Mintzberg said that each manager performs ten roles classified into three categories these are interpersonal roles, informational roles, and the decisional roles. Table (2-9) illustrates the roles and shows how each information system can help managers performing the ten roles that Mintzberg clarified. The table shows that in performing some of the managers' roles, managers should depend on their own experience and judgement, whilst in the rest they can utilize different IS.

[8] Adapted from (Turbanand Aronson, 1998; Alter,1992).

Category	Role	Activity	Relevant IS
Interpersonal	-Figurehead	: Performing ceremonial duties such as greeting visitors.	-
	-Leader	: Direct and motivate subordinates.	-
	-Liaison	: Maintain information links inside and outside the organisation.	OAS and EIS
Informational	-Monitor	: Seek information and scan reports.	MIS
	-Disseminator	: Forward information to others in the organisation.	OAS
	-Spokesman	: Transmit information to outsider through speeches and reports.	EIS
Decisional	-Entrepreneur	: Initiate improvement projects.	-
	-Disturbance handler	: Corrective actions in crises and conflicts.	MIS
	-Resource allocator	: Decide who gets resources and when	DSS
	-Negotiator	: With concern to sales, purchasing, unions, and budgets.	-

Table (2-9). The managers' roles and IS[9].

Chapter summary

- Data are raw facts.
- Information is processed data, that must be clear, on time, complete, correct, relevant, and at the required frequency.
- Knowledge is a combination of experience, accumulated learning, and information that have been organised and analyzed to be understandable and applicable to a specific decision situation.
- O'Brien (1996) differentiated between the various IS based on their usage on by managers at the different management levels. Figure (2-5) illustrates this differentiation.

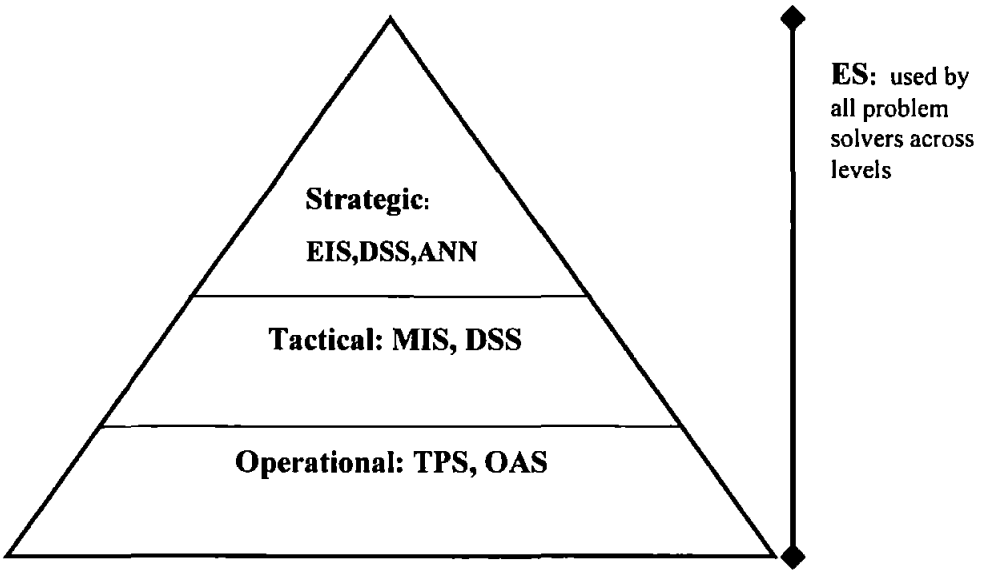


Figure (2-5). Management levels and the various IS[10].

- An information system (IS) is a combination of work practices, information, people, and information technologies organised to accomplish organisational goals.
- Managers use information systems to assist them taking better decision. For the structured decisions managers use TPS and MIS. For the unstructured decisions, DSS, ES and ANN are used. EIS are special type of information systems that support unstructured decisions.
- IS are helping organisations achieve competitive advantages. The strategic role of IS involves the development of products and services and capabilities that give the organisation strategic advantages over the strategic forces in the market, the information system that does this function is *called strategic information system* (SIS). SIS can be any kind of IS, it might be TPS, MIS, EIS, or DSS.
- TPS record and collects data about the daily transactions taking place in any organisation; they are short-term in nature and have no decision capability.
- MIS provide the information required for managing the organisation. MIS require TPS as a prerequisite.
- ES are designed to help managers make better decisions in certain areas. They are interactive CBIS that respond to questions and give recommendations.
- OAS refer to all the CBIS associated with general office work applications.
- ANN attempt to tease out meaningful patters from vast amounts of data and can recognize too many patterns.

[9] Adapted from (Mintzberg, 1971).

- EIS/ESS are highly interactive systems providing managers and executives with flexible access to information for monitoring operating results and general business conditions. Executives use EIS without any aid from intermediaries.
- Decisions are taken by either individual manager or group of them, so when the decision is to be taken by a group this is where we have GDSS.
- DSS components are data management subsystem, model management subsystem, knowledge management subsystem, user interface subsystem and the system user.
- Hardware affects the functionality and capabilities of the DSS.
- DSS can be classified according to many factors including the degree of action implication of system outputs, or in other words the extent that the system outputs could directly support the decision, the DSS orientation, type of problems, the degree of non-procedurality, and the type of DSS support.
- DSS can be developed using a number of approaches including the SDLC, prototyping or end-user.
- The combination of DSS, DW, and KDD constitutes a new approach in DSS development. The DW adds the strategic value to the DSS through the wealth of information available on it that contains the organisation's history across years. The KDD as a front-end tool to the DW can extract valuable information and patterns and unknown facts from the DW and present them to the DSS user for the goal of achieving better quality of decisions. DW is discussed in details in chapter three, while the KDD is discussed in chapter four. Moreover, chapter five will introduce this blend of technologies DSS, DW, and KDD to formulate the proposed DSS methodology.

[10] Adapted from (O'Brien, 1997).

Chapter three

Data Warehousing

This chapter covers the details of the data warehouse (DW). The chapter starts with introducing the data needs of organisations and then investigates the different data sources in organisations; internal, external, archival and personal that are required to meet these needs. The relationship between these data sources and the data warehouse is developed. The traditional database models; relational, hierarchical, network, object-oriented, and the multi-media are then discussed because these models are used to develop operational data stores (ODS). Definitions and features of the data warehouse are illustrated and a recommended DW definition is proposed. The characteristics of a data warehouse including time-variant, non-volatile, subject-oriented and integrated are discussed together with the benefits both tangible and intangible. The differences between data warehouses, data marts, and enterprise data warehouses are discussed. The star schema structure is introduced in detail. The data warehouse components are discussed followed by the Client/Server concept and its relationship with data warehousing. Data warehouse development approaches, which are top-down and bottom-up are critically evaluated followed by defining the users of the data warehouse. Then, the DW development strategy is discussed. The chapter ends by suggesting some DW development guidelines.

3-1 Organisational needs

(Srivastava and Chen 1999; Barquin, 1997; Paller, 1997) have stated that in order to survive and succeed in today's global environment organisations' needs from data have become more variable because of the following:

1. Decisions need to be taken quickly and correctly using all the available data;
2. Users are not computer professionals so they need all the relevant data concerning a specific business problem to be stored in one place (Berson and Smith, 1997);
3. The amount of data concerning a specific business problem is increasing;

4. It is becoming increasingly important to be able to obtain a comprehensive and integrated view of the enterprise for the purpose of making decisions about the business across time periods;
5. Decisions sometimes require historical analysis e.g. can we offer certain clients a promotional offer based on their past history? What sort of offers work best for which type of clients?
6. Increasingly businesses are working closer together and are able to share and exchange data in what is known as a *strategic alliance*. A strategic alliance is a mutually dependent relationship where the success or failure of one party affects the other (Reynolds, 1995; Coyle, et al., 1992). Examples are:

   Allelix (www.allelix.com/alliance.htm);

   ViaNet (www.vianetcorp.com/solutions/alliance.htm);

   WillCam (www.willcam.com/assoc.htm);

   and **Rover Group** (www.rover.co.uk)
7. Identifying trends in the business (Onder and Nash, 1999).

3-2 Data sources

To respond to these needs various departments in organisations store data about internal transactions and about their external environment. Further organisations need to store these data for a number of years in a historical *(archival)* database to identify patterns and/or to meet legislative requirements e.g. tax regulations. These data sources are then accessed by decision makers to reach a better business understanding and improve the quality of their decisions. Sometimes decision makers keep their own experience in a separate database (Turban and Aronson, 1998) These different types of data are illustrated bellow.

3-2-1 Internal data

These are the data sources of an organisation that cover the whole business, e.g. data about employees, daily transactions, products, stock levels, customers etc. Usually internal data

is stored in one or more databases. For example, the organisation employees' data may be stored in the personnel database, the product details stored in the products database, and the customer details in the customer database.

Internal data sources are also created by organisations when using TPS, because these systems store data about business transactions. All these internal data sources are sometimes referred to as operational data sources (ODS) (Hadden, 1998a) or on line transaction processing systems (OLTP) (Berson and Smith, 1997; Devlin, 1997);

3-2-2 External data

This is data that comes to the organisation from outside sources. There are many types e.g. government reports, federal publications, research institutions, commercial data banks, access to suppliers and customers databases, and the Internet (Turban and Aronson, 1998). External data sources are used in EIS and DSS to enhance the strategic and long-term decisions. Examples for the commercial data banks include CompuServe, Compustat, Data Stream, Dow Jones Information service, and Lockheed Information systems;

3-2-3 Archival or historical data

When an organisation needs to store data about a specific topic for several years, it uses an archival or historical database. The archival database can contain either internal or external data sources or both;

3-2-4 Personal data

This data source includes the manager's own experience and opinions and/or estimates about market share, additional customer data or other policies. These personal data sources are used in EIS and DSS.

3-3 Database models

At the operational level data is collected into a number of separate data sources which will form the basis of the ODS and the data warehouse (DW).

For organisations to be able to develop operational data stores from their different data sources there are various database models that can be used. There are three fundamental database (DB) models: relational, hierarchical, and network. There are also new DB models e.g. object-oriented, multi-media, and the star schema structure (Elmasri and Navathe, 2000; Livingston and Rumsby, 1997; Date, 1995; Pratt and Adamski, 1987). These different types of database models are discussed in the following sections.

3-3-1 The relational model

The original concept for the relational model was proposed by Dr.Codd in the 1970's. This model is the most frequently used database model within the DBMS context (Elmasri and Navathe, 2000; Abiteoul, 1995; Date, 1995). It is also the dominant database model in DSS applications, and frequently used in the development of a DW (Berson and Smith, 1997). It allows the user to think of the DB model in terms of two-dimensional tables. A table consists of rows and columns, rows are the data records and columns are the individual fields. A group of related rows/records constitute one table and a group of related tables constitutes a DB. Data tables are joined to each others be creating *relationships*. Relationships have *cardinality ratios*, these ratios are: one-to-one (1:1), one-to-many (1:M), many-to-one (M:1), or many-to-many (M:N). Many-to-many relationships are structured as two one-to-many relationships. Relationships also have *degrees*, these degrees are: recursive (one table and itself), binary (two tables), ternay (three), and n-ary (more than three). Each table has a unique identifier (i.e. *primary key*), and if two tables are participating in a relationship, they require a common field of the same type and size. That filed is the primary key of the independent table that migrates to the dependent table (i.e. *foreign key*).

The advantages of the relational model are:

- Easy to use;
- Flexible in design;
- Supports multiple access queries;

- Provides DB-Application independence (i.e. DB structure can be changed without having to change the applications);
- Easily expanded;
- Easy to manipulate the data using SQL;
- Able to eliminate redundant data through the *normalization* process. The normalization process has been divided into a number of steps called normal forms. Normalization also enhances the response time of the DB (Elmasri and Navathe, 2000; Abiteboul, 1995). A typical relational database model is shown below.

Student**

ID*	Name	Age	Code
100	Ahmed	20	11
200	Ali	18	11
300	Mai	21	22

Course

Course SNO*	Title
BUS01	Accounting
BUS02	Management
COM11	PASCAL

Major**

CODE*	Name
11	Computer
22	Economics

Registration***

Ref No*	ID	Course SNO
1	300	BUS01
2	200	COM11
3	300	BUS02

* Primary key fields.
** Major and Student have a 1:M binary relationship.
*** Student and Course have a M:N binary relationship broken down using Registration.

Figure (3-1). The Relational model.

The software package that is used to create and maintain a relational database is called a relational database management system (RDBMS) e.g. Oracle, Sybase, IBM, MS SQL, and Informix. Since the relational model concept was proposed by Codd, he identified 12 rules that any RDBMS should meet. In fact not all of Codd's 12 rules are met in all RDBMS, in practice products are considered RDBMS even if they do not strictly meet the 12 rules (Codd, 1990). The 12 rules are (*Data in tables, data is logically accessible, nulls are treated as unknown, DB is self-describing, DBMS uses single language to communicate with, data viewing alternatives, supports set-based operations, physical data*

independence, logical data independence, data integrity, supports distributed operations, and data integrity cannot be subverted). More details can be found on (Codd, 1990).

3-3-2 Hierarchical

The hierarchical model uses two main data structures; *records* and *parent-child relationships* (PCR). A record is a group of related values for data items (i.e. fields or attributes). Records of the same type are grouped together in a record type (e.g. product, emplyee, course..etc). A PCR is a 1:N relationship between two record types; whereby the record type on the 1-side represents the parent and the record type on the N-side represents the child. A PCR has many occurences, each occurrence consists of one record of the parent record type and zero or more record of the child record type (Elmasri and Navathe, 2000).

In the hierarchical model, a DB consists of a number of *hierarchical schemas*, each hierarchical schema has record types and PCRs. A hierarchical DB schemas stores the data in a top-down order which corresponds to a *tree data structure*; on which the record types represent the *nodes* whilst the PCRs represent the *edges*, and there is always one path between any two nodes (Parsaye, et al., 1989). The basic operation in a hierarchy is the tree search, when a query is processed the nodes that meet the conditions of the query will be returned. Data has to be linked in hierarchies and each node has only one parent but in real world applications this condition rarely holds.

M:N relationships can not be represented directly by the hierarchical model, because every PCR is a 1:N relationship and a record type can not participate as a child in more than one PCR. The M:N is represented by *duplicating* the child record type instances. For example if the relationship between Student and course is M:N, the same student record appears under every course which the student books. This treatment to the M:N relationships increases both the size of the DB and the response time. The way this model handles the M:N relationships is one of its disadvantages which negatively affected the model usage.

The hierarchical model has poor flexibility, it is hard to maintain and difficult to change its design (Parsaye, et al., 1989). The hierarchical model continues to be used even though many DB applications are migrating to the relational DB model. This is because classical applications were written using a hierarchical model e.g. some accounting and payroll systems.

An example is shown below in figure (3-2). In a hierarchical schema consists of both record types (e.g. College, Major, Department) and PCR (e.g. [College, Major], [Major, Student]) and must follow the following roles:

1. The *root* record type (i.e. College) does not participate as a child in any PCR;
2. Every record type except the root can participate as a child record type in exactly one PCR type;
3. A record type can participate as parent record type in zero to many PCR types;
4. When a record type does not participate as parent record type, it is then called a *leaf* of the hierarchical schema;
5. When a record type participates as parent in many PCR, its child record types are ordered. That order is displayed from left to right in the hierarchical schema.

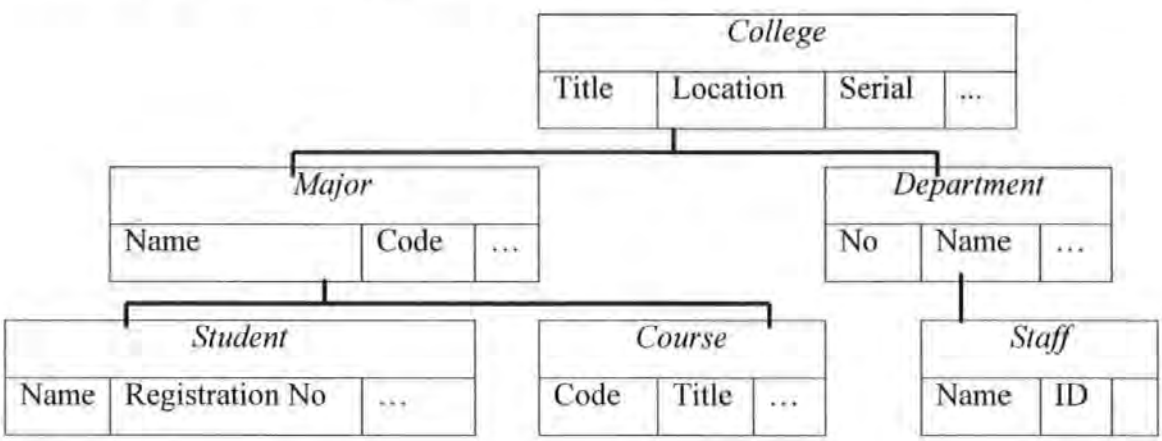


Figure (3-2). Hierarchical model[1].

There are fundamental differences between the hierarchical model and the relational model. In a relational model there is no hierarchy among the tables and any table can be accessed directly and can potentially linked with any other table. The hierarchical model is

[1] Adapted from (Elmasri and Navathe, 2000).

hard to be coded, the design is very difficult to expand or modify, changes typically require huge coding time and efforts. Hence, the hierarchical model is inflexible.

3-3-3 Network

The original network model was presented in the CODASYL Data Base Task Group's report in 1971, this is why this model is sometimes called the DBTG model or the CODASYL (Elmasri and Navathe, 2000). The network model has two basic data structures; *records*, and *sets*. Data is stored in records; each record consists of a group of related data items (i.e. attributes or fields). Records are classified into *record types*; each record type describes the same structure and stores the same type of information. Student, department, major are examples of record types. A set type is a 1: N relationship between two record types, this constraint is always by the DBMS in the network model. To represent 1:1 relationships an application program must be written to enforce this constraint. M:N relationships between two record types can not be represented by a single set type, a new linking record type is to be added. These restrictions on the representation of the 1:1 and M:N relationships increased the use of the relational model, and has significantly reduced the use of the network model.

The network model saves storage space through the sharing of data items. The network model uses additional pointers to give more flexibility than the hierarchical model. The network model allows more complex links between nodes. Moreover, the network model permit the child or leaf nodes to have any number of parents (including zero), whilst in the hierarchical model a leaf has only one parent. Thus the hierarchical model may be viewed as a special case of the network model where each node is linked to a parent node only (Parsaye, et al., 1989).

Figure (3-3) depicts a network model example. To represent a relationship (i.e. set type) three elements need to be identified; name for the set type (*Major_Student*), owner record type (*Major*), and a member record type (*Student*). Many set *occurrences* (or instances)

correspond to a set type, each occurrence contains one record from the owner record type and set of of records from the member record type. Any record from the member record type can not exist in more than one set occurrence (i.e. 1:M constraint).

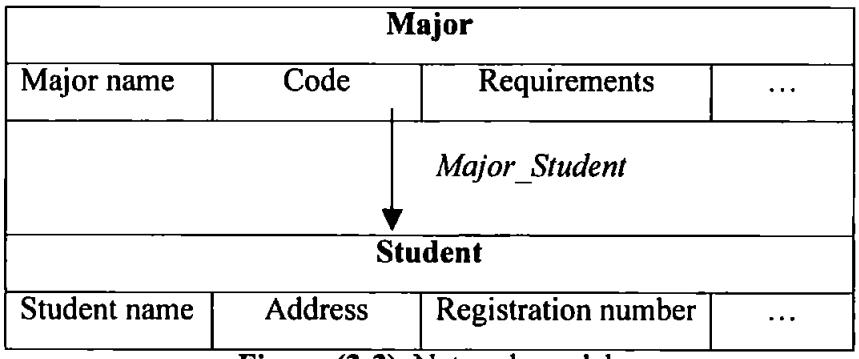


Figure (3-3). Network model.

The relational model is able to provide higher levels of flexibility than the network model, in addition to the relational model's ease of use. On the other hand, relationships are very complex in the network model and are very hard to maintain and/or implement. According to a technical report by David R. Frick & Company (2001), the network model is for all practical purposes obsolete.

3-3-4 Object-oriented

This model is used with complex applications which require accessibility to data that have complex and inter-related relationships (Date, 1995). For example computer aided design and manufacturing (CAD/CAM), computer integrated manufacturing (CIM), and geographic information systems (GIS). Relational, hierarchical, or network data models can not support these applications efficiently. The software package that utilizes the object oriented model in creating and maintaining databases is called object-oriented database management systems (OODBMS). OODBMS are based on the object-oriented programming (OOP). OODBMS allow the analysis of the DB in terms of objects. Abstraction is used to develop the inheritance between object levels, encapsulation allows the DB designer to store conventional and procedural code within the same objects. OODBMS deal with data as objects that have relevant structure and behavior; they use the

concept of class and subclass to enable the inheritance concept implementation. OODBMS have the power to be used in many complex management support applications like DSS.

The object-oriented model and the network model are both navigational in nature, however, the data structure capabilities of the network model are much more elaborative. On the other hand, the network model lacks some of the desirable features of the object-oriented model such as inheritance and encapsulation (Elmasri and Navathe, 2000).

3-3-5 Multi-media

The multi-media model manages data in many formats; text, numeric, images, bit-maps, pictures, hypertext, video clips, sounds and multi-dimensional images (virtual reality). The software that utilizes the multi-media model in creating and maintaining databases is called a multi-media database management system (MMDBMS). It is possible for the other DBMS to be able to deal with data in the previous formats. For example, ORACLE, SYBASE, and INFORMIX store these data types under what they term binary large objects (BLOBS).

3-4 Data Warehouse (DW)

A data warehouse is designed specifically for decision support queries therefore only data that is needed for decision support is extracted from the operational data sources and stored in the warehouse. Designing a data warehouse requires special knowledge because the data model must contain only the data needed by users. Including data that is not required will reduce the speed of access. In order to meet these requirements the database model used for the design of the warehouse will be different from that used to model an operational database. After creating a corporate data model for the data warehouse the design of a specific data management environment has to be performed. If the operational data is contained in a number of databases the relevant data will need to be copied from these databases to the data warehouse.

Even when the structure of a data warehouse is complete, it will need to evolve over time to meet the changing needs of the business and its environment. For example *if modifications have been carried out to the attributes of the operational database, this can change the data model of the data warehouse. These changes can be carried out with the aid of appropriate data management tools that can control the DW environment.* Once the data warehouse is set up, snapshots of this data can be taken and stored on local database servers as necessary.

Several DW definitions have been proposed by Inmon & Hackathorn, Widom, Berson, Mattison, Barquin, Berson and Smith, Devlin, Adamson and Venerable, & Turban and Aronson. A discussion of these definitions follows.

3-4-1 Inmon and Hackathorn (1994)

Inmon and Hackathorn defined the data warehouse as being a subject-oriented, integrated, time-variant, and non volatile collection of data that is used in support of management's decision making process.

3-4-2 Widom (1995)

Widom defined the DW as architecture for bringing together selected data from multiple databases or other information sources into a single repository called the data warehouse, suitable for direct querying or analysis.

3-4-3 Berson (1996)

Berson defined the data warehouse as the means for strategic data usage. Berson also said that a DW can be viewed as a foundation for an IS that owns the following features:

- It is used extensively for READ type operations;
- It includes large volumes of records in a few number of tables;
- Each query is processed in large data sets using multi-join tables;
- It contains current and historical data;
- It allows the storage of metadata that contains data summaries, which makes the search process easier;

- It is periodically updated;
- It supports a small number of users;
- It is a database designed for analytical tasks;
- It uses data from different databases and from various applications.

3-4-4 Kimball (1996)

Kimball defined the datawarehouse as a place where people are able to access their data from. Kimball defined the data warehouse goals as follows:

- Providing access to corporate or organisational data;
- The data stored in the data warehouse is consistent (i.e. different people have the same number when they inquire on the same item);
- Enabling the drilling down and slicing up capabilities (i.e. more details or aggregations);
- The warehouse is not only a data storage but provides query, analyze, and information presentation tools;
- The data warehouse is the place where used data is published;
- The quality of the data in the data warehouse is a driver of business reengineering.

3-4-5 Mattison (1997)

Mattison said that a data warehouse is a collection of data copied from other systems and assembled in one place. Once it is assembled it becomes available to end-users who can use it to support different kinds of business decision support systems and information activities.

3-4-6 Barquin (1997)

Barquin defined the DW as a process through which organisations extract value from their informational assets.

3-4-7 Berson and Smith (1997)

Berson and Smith asserted that a data warehouse is not a product rather it is an environment. They have defined the DW as a blend of technologies and components the

aimed at the integration of operational databases into an environment that permits the strategic use of data.

3-4-8 Devlin (1997)

Devlin said that the DW is simply a single, complete, and consistent store of data obtained from a various sources and made available to end users in a way they can understand and use in a business context.

3-4-9 Adamson and Venerable (1998)

Adamson and Venerable said that while most computer systems are designed to capture and store data, the DW is designed for getting data out. It is all about getting answers to business questions.

3-4-10 Turban and Aronson (1998)

Turban and Aronson explained that the purpose of the DW is to establish a data repository that prepares the operational databases in an organisation in an accessible and ready-to-use format for DSS and EIS. Only the data that is required for DSS or EIS is extracted from the operational databases and then stored in the DW. Data warehousing or information warehousing as it is sometimes called combines data from different sources into one for end-user access.

3-4-11 Summary

The following table (3-1) summarizes the previous defintions each with its focus point(s).

Reseacher	Focus
Inmon and Hackathorn 94	Subject-oriented Integrated Time-variant Volatile
Widom 95	Multiple data sources
Berson 96	Strategic use Features
Kimball 96	Place for data access Goals
Mattison 97	Data sources End-user
Barquin 97	Information value
Berson and Smith 97	Components
Devlin 97	Data sources End user
Adamson and Venerable 98	Data sources
Turban and Aronson 98	Data sources Front-end tools

Table (3-1). DW definitions' focus.

Each of the definitions tries to define the DW taking into consideration certain point of view. Some definitions are technical (Berson and Smith, 1997), others are about the use of DW (Berson, 1996), the features (Inmon and Hackatorn, 1994), or the goals (Kimball, 1996). However, no definition has been found to be comprehensive covering the data sources, front-end, and the purpose of the DW in a business context. This is why the definition that will be adopted by this research is found in the next section.

3-4-12 A DW definition

The DW working definition that will be used in this research: "*A DW is a group of data extracted from different sources; internal, external, historical, and personal data archived in one or more data stores. The purpose of constructing a DW is to provide the DSS and the decision maker with the necessary data, which when transformed into information, will provide a better understanding of the business problem.*"

3-5 Data Warehouse Characteristics

According to (Inmon, 1993) there are four characteristics that generally describe a DW:

1. *Time-variant.* The DW contains data gathered from different periods. The DW contains a place for storing historical data that can be used for comparisons, trends, or forecasting. Historical data can be over twenty years old;
2. *Non-volatile.* The objective of using the DW is to respond to management requests for information. This data is extracted from the operational database and then loaded into the DW database. This means that a data warehouse will always be filled with historical data and should be updated regularly from the operational database. Some DW components are *static* that is they contain data that does not change over time like a country's past history or events. Whilst another DW components are automatically updated from their sources and they are called *active DW* components;
3. *Subject-oriented.* Data are organized according to subject instead of application. Examples of subjects are marketing, production, personnel, sales etc. The data are organized by subjects that are relevant to the decision support systems;
4. *Integrated.* In many organisations the same piece of data may exist on several databases, to overcome the data redundancy problem there has to be an integration of data sources to avoid duplication.

3-6 DW benefits

A data warehouse increases the decision maker's productivity by providing accessible data in a ready to use format (Wixom and Watson, 2001; Turban and Aronson, 1998). In a situation where an organisation does not have a DW, the user queries are derived from the operational TPS and/or if the queries are complex and deal with thousands or millions of records this will downgrade the performance of the TPS. In the case where an organisation has a DW, the DW isolates the operational databases from the query processing of the DSS or EIS. Hence this does not affect the ODS performance.

A DW stores the internal, external, historical, and personal data that are of interest to business managers into a single consolidated system which reduces the time that these managers need to spend to find and analyze data.

A DW also eliminates the need for user managers to have computing expertise to enable them to navigate through different databases and extract the relevant data required to investigate their current problem. The benefits of the DW can be classified as tangible or intangible benefits (Berson and Smith, 1997).

3-6-1 DW tangible benefits

The following tangible benefits have been reported:

- Product inventory turnover is improved;
- More cost-effective decision making process by separating the query processing from the operational databases (Wixom and Watson, 2001);
- Enhancing asset and liability management by providing the overall picture of the enterprise purchasing and inventory transactions;
- Supporting the corporate strategy that positions the clients at the center of all operations. This client-centered strategy could not be achieved without a DW (Cooper et. al, 1999);
- To record the past history accurately;
- Supporting the Reengineering of decisional processes (Humphries, et al., 1999).

3-6-2 DW intangible benefits

A study by Onder and Nash (1999) reported the following intangible benefits:

- Improving productivity by keeping all the required data in a single location;
- Reduces redundant processing, support, and software to enhance DSS applications;
- Enhancing the work process, which also affects the success of business process reengineering;
- Improve customer service;
- Organisations will be able to exceed competitor capabilities and achieve competitive advantages.

3-7 Data marts, data warehouses, and enterprise data warehouses

In the next sections the differences between data marts and data warehouses will be elaborated, the enterprise data warehouse concept will be introduced, and the relationship between the three terms will be evaluated.

3-7-1 Data marts and data warehouses

Some scholars referred to the two terms *data warehouse* and *data mart* as synonyms, whilst others differentiated between the two terms. Hence, the previous studies are classified into two categories; the first includes those who set differences between the two terms, whilst the second category includes those who did not distinguish between the two terms. *The two categories are analyzed as follows*:

1. Humphries, et al., 1999; Srivastava and Chen, 1999; Sperley, 1999; Adamson and Venerable, 1998; Hadden, 1998a; Paller, 97; Levin, 1997; Adriaans and Zantinge, 1996 distinguished between DW and data marts. However, the criteria for differentiation varied amongst the scholars.
    a. Sperley (1999: 15) said, "another structure, *the data mart*, closely resembles a data warehouse. In much the way that a "*food mart*" has less selection and availability than those of a supermarket, a data mart is a miniature data

warehouse. A data mart typically has a smaller number of data, fewer subject areas, and less history. A data mart can be thought of as a logically or physically partitioned subset of a data warehouse. A data mart is usually constructed to serve the needs of a particular user community.";

b. Paller (1997) defined the data mart as a departmental data warehouse;

c. Levin (1997: 70) said, "custom developed applications may be able to access data warehouse directly, while proprietary package solutions often need to extract data from the data warehouse and import the data into its own data stores. Some of these types of applications overlap with multi-dimensional analysis types. If separate data stores are built, these applications may be termed data marts". That is, Levin defined the data mart as a copy of the DW that is accessed and stored by some applications' users, however if this data will to be accessed directly from the DW and not stored on a local machine, the data mart does not exist;

d. Humphries, et al., 1999; Srivastava and Chen, 1999; Adamson and Venerable, 1998; Hadden, 1998a; Adriaans and Zantinge, 1996 defined the data mart as a local data warehouse that is classified by subject e.g. the marketing data mart, and the personnel data mart.

2. Sørensen and Alnor, 1999; Edlestein, 1997; Kimball, 1996 did not distinguish between the two terms. Actually the term "data mart" was not mentioned neither in Sørensen and Alnor nor in Kimball's work. However, Edlestein (1997: 37) said, "data marts are data warehouses in their own right and may be as large (or larger) than the data warehouse that spawned them."

For the purpose of this thesis:

1. The data mart is different from the DW. This distinction is supported by the previous studies of Humphries, et al., 1999; Srivastava and Chen, 1999; Sperley, 1999;

Adamson and Venerable, 1998; Hadden, 1998a; Paller, 1997; Levin, 1997; Adriaans and Zantinge, 1996;

2. Sperley's definition for data mart will be adopted. *The reasons for adopting Sperley's definition are*:
    a. The definition is comprehensive and more detailed than the other definitions proposed by Humphries, et al., 1999; Srivastava and Chen, 1999; Adamson and Venerable, 1998; Hadden, 1998a; Levin, 1997; Paller, 1997; Adriaans and Zantinge, 1996;
    b. According to the definition a data mart is:
        i. Smaller than a DW in terms of number of records and fields, this means that the response time of the data mart will be faster than the DW for the same user on the same query;
        ii. A logical and/or physical subset of the DW. Logical subset means it can be part of the original DW data model, whilst physical subset means it can be stored in a local machine. This means that more than one data mart could be built based on the same DW;
        iii. Constructed to serve the needs of a particular user community. This means that many users are able to share the data mart;
        iv. Data marts and DW can co-exist.

3-7-2 Enterprise data warehouse (EDW)

Adamson and Venerable (1998: 463) said, "*Enterprise data warehouse* is a planned, integrated, managed store of relevant corporate data optimized for analysis, query, and reporting functions". An EDW contains large number of fields and millions of data records about the entire organisation. For organisations to build EDW, the following alternatives have been used:

1. Build DW and/or data marts for selected departments without building the EDW;

2. Build DW and/or data marts for all departments, and then construct the EDW from these data marts;
3. Build the EDW directly without building separate DW and/or data marts.

The choice of these alternatives depends on the budget devoted for the EDW project and on the business requirements. Figure (3-4) illustrates the relationship between data marts, DW, and EDW.

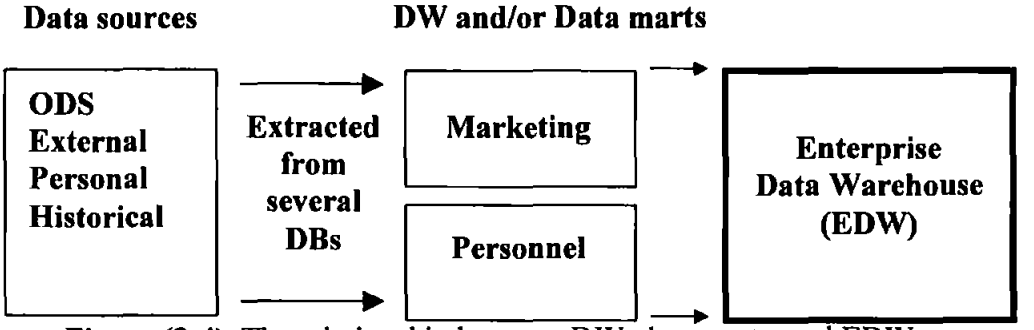


Figure (3-4). The relationship between DW, data marts, and EDW.

3-8 The difference between the ODS and DW

Organisations use TPS to store their business transactions. These TPS have a certain database design requirements and for this reason these types of databases are known as the *operational data stores (ODS) or on-line transaction processing applications (OLTP).* They are optimized to handle large numbers of concurrent users either storing/editing transactions or processing queries. Whilst thousands of users might be connected to the same ODS performing millions of transactions, in the DW there are typically a few users connected to the DW primarily for analysis and complex query purposes (Edelstein, 1997). Deriving information online from an ODS may be self-defeating because it can dramatically downgrade the performance of the operational database systems. It is also time and effort consuming and generally requires programming expertise beyond the capability of most users.

A study by (Barquin, 1997) showed that many managers were disappointed to find that they have major media compatibility problems and also that they are unable to recover

historical data. From the users' perspective a DW is capable of resolving the technology problems associated with incompatibility between ODS by introducing one central repository (Humphries, et. al, 1999; Berson and Smith, 1997).

Data is typically loaded from the ODS systems in batch formats. When the data is loaded from the ODS into the DW the data must be checked and cleaned from noise and other sources of errors like incorrectness or incompleteness (Livingston and Rumsby, 1997). As the data is in the DW nothing is done to manage or check its quality.

Once a DW has been developed the historical data can be used to alert the managers to business problems. This issue is related to the fact that, in general, the earlier the source of the problem can be identified the lower the cost of fixing it. With the trend towards organisations sharing data, a DW provides an efficient method of supporting these strategic alliances. An *inter-organisation data warehouse* can be established between two organisations where each one makes a DW available to the other one for the purpose of facilitating business transactions (Devlin, 1997). In the area of decision support and business analysis organisations find their needs supported by DW. The following table (3-2) illustrates the differences between the DW and ODS.

Criteria	ODS	DW
-Purpose	Transaction needs	Strategic needs
-Clients	Users, Administrators	Executives, Managers
-Systems type	Batch	DSS, EIS
-Content	Current values	Summaries, Subsets
-Data actions	Create, Read, Update, Delete, Print	Read-Only
-Data sources	Internal	Internal, Archival, External & Personal
-Size in bytes	Small (MB to GB)	Large (GB to TB)
-Orientation	Application	Strategic
-Response time	Fast	Slow
-Integration	Partially	Fully

Table (3-2). ODS Vs DW[2].

[2] Adapted from (Hadden, 1998a).

Through the incorporation of four different data sources (i.e. *internal*, *external*, *archival*, and *personal*) into the DW, the DW enlarges the data available to managers and executives for the purpose of better decision quality (Hadden, 1998a).

3-9 The Star Schema structure

3-9-1 Overview

There is a primary difference between a database that is designed for operational systems (e.g. stock levels system, sales system, and payroll systems) and the database design for the data warehouse. The ODS databases provide the DW with a source of data, however, they lack the functions required to perform efficient analysis and produce reliable results that decision makers really need (Livingston and Rumsby, 1997). The contents of the DW are relatively stable whilst the contents of the ODS change as each transaction is initiated.

The best way to build the DW database is by using the star schema structure (sometimes referred to *as multi-dimensional data modelling*-MDDM). The Star schema or MDDM captures the *measurements* of importance to the business and the *parameters* by which the business measurements are broken down. It is a direct reflection on how business processes happen. The measurements are referred to as FACTS, whilst the parameters by which a measurement can be viewed are called DIMENSIONS (Sørensen and Alnor, 1999; Firestone, 1998; Adamson and Venerable, 1998; Kimball, 1996).

A simple star consists of group of tables that describe the dimensions of the business arranged around a central table that contains the business facts. The smaller outer tables are the points of the star, the larger table in the center is the star from which the points radiate. The star schema relies on two major components the *facts* and the *dimensions*. Sometimes for the purpose of enhancing the performance of the DW summary tables are created. Further, to transform the ODS database model to a star schema structure a de-normalization process takes place. Indexing is another technique by which the DW

performance can be leveraged. A discussion of these components and techniques follows in the following sections.

3-9-2 Fact tables

This is the central table and generally it is the biggest table in the DW database in terms of records. The DW DB can contain one or more fact tables. When more than one fact table exist, they are called a *fact table family* or sometimes called a *multi-star structure* (Sørensen and Alnor, 1999; Humphries, 1999; Adamson and Venerable, 1998; Livingston and Rumsby, 1997; Kimball, 1996). Examples of fact tables are sales, orders, budgets, shipments, students, and accounts. An important factor in designing the fact table fields is to make them as small as possible in terms of the data size. This is because the size of the fact table will grow dramatically and frequently stores millions of records. Fact tables are built with a *multi part primary key* (sometimes called *concatenated* or *composite* key), with the key typically consisting of more than one field. Each field points to a matching field in a dimension table. Through these links referential integrity is enforced.

Besides the key of the fact table there are other attributes, those attributes are called *numerical measurements of the business*. Examples are total sales, quantity received, average age, and student GPA. In a few business occasions the fact table contains only the primary key attributes, in which cases it is called a *factless fact table*.

Kimball (1996) said that the most useful fact table should be *additive*. The reason for that is, every query runs against the fact table is expected to work on thousands or millions of records that require summarization and aggregations which is very hard to achieve if the fact table attributes are *non-additive* (i.e. non-additive attributes cannot be added at all). Additionally, some fact tables are *semi-additive*, that is some of the attributes are additive, and the others are non-additive. For the semi-additive tables other functions could work. E.g. count (instead of sum), or they can be added along the dimensions.

3-9-3 Dimension tables

Dimension tables are the points of the star or the fact table. Examples of dimension tables are time, markets, products, courses, staff members, majors, and vendors. Dimension tables use both character (i.e textual) and numeric data types so their fields are usually much bigger in size than the fact table fields. The number of rows in the dimension table is less than the number of rows in the fact table. Typically, the dimension table contains single-part primary key. Each dimension table has a fixed number of records, for example the: list of courses; list of products; list of employees; or list of markets.

3-9-4 The TIME Dimension

Since the DW includes offloading archival/historical data from the operational systems, so each fact in the DW must be time-stamped (Humphries, et al., 1999). This requires each DW and/or data mart to include a TIME dimension in the design (Firestone, 1998; Hadden, 1998a).

Kimball (1996) said that the time dimension is virtually guaranteed to be part of each DW because every data warehouse is a time series. However, if the business uses the dates and time spans on a year or month basis, then in this situation the time dimension table could be deleted.

3-9-5 The Granularity of the Fact table

The term *granularity* (sometimes referred to as *grain of the fact table*) describes the level of detail stored in the fact table and follows the level of detail of its related dimensions (Humphries, et. al, 1999; Adamson and Venerable, 1998; Hadden, 1998a; Kimball, 1996). Determining the grain of the fact table is a very critical decision. Granularity at too high level prevents the users of the DW from drilling down into further details of the data. Granularity at a low level of detail results in an enormous increases in the DW size and consequently affects both cost and performance.

For example, if the record in the time dimension table represents a semester, the record in the Colleges dimension table represents a College, the record in the Majors dimension

table represents a Major, and the record in the Students_GPA fact table represents a student then the grain of the fact table in connection to these dimensions would be: the student GPA per College per Major per Semester.

According to Kimball (1996), it is important to realize that the decision on the level of details to be included in the fact table (i.e. granularity) should be taken before the design of the fact and dimension tables (i.e. attributes).

It is worth mentioning here that Kimball (1996), recommended the following steps in designing a DW:

1. *choose the business process to model*;
2. *decide on the granularity of the data warehouse*;
3. *design the dimension tables*;
4. *design the fact table(s).*

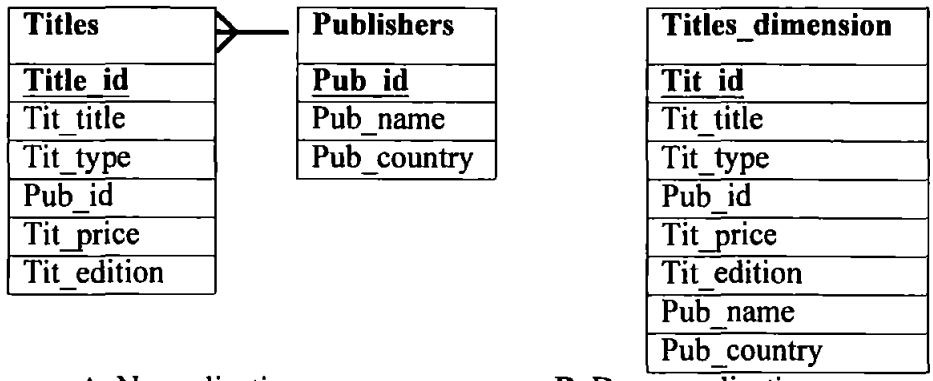
3-9-6 Summary tables

A summary table is a DW table that includes data frequently retrieved by users. Instead of searching in the entire fact table a snapshot is taken and stored in a summary table, when the user invokes the relevant query the result comes from the summary table (Hadden, 1998a). Hence, summary tables allow the DW to respond rapidly to known or anticipated business queries. A survey by the Compaq Corporation (1999) to examine MS-SQL Server data marts tools showed that it is not recommended to create summary tables if the query is used infrequently or if the data retrieved by this query is more than 20% of the rows in the fact table. In these cases it is more efficient to retrieve these data from the fact table directly.

3-9-7 De-normalization

Unlike the relational model, the star schema structure uses the *de-normalization* process to enhance the performance of the DB tables. De-normalization helps to reduce the number of joins between the tables, thus making the query writing process easier, and also reduces the query execution time (Sørensen and Alnor, 1999; Hadden, 1998a). Whereas the

normalization process tries to split-up tables, the de-normalization process rolls-up all the data about the dimension in one table. Figure (3-5) shows the difference.



A. Normalization **B.** De-normalization

Figure (3-5). Normalization vs. De-normalization[3].

3-9-8 Indexing

Since the DW is built to support the strategic use of data the DSS/ EIS will often use queries that require extensive amounts of processing time to extract the required information. In order to reduce the processing time the DW designer makes extensive use of indexing.

Users have come to expect fast response to any computer based question. However, as the volume of data stored grows and the usage expands, the DW will not able to respond quickly to all users' queries. The index used in the DW is big factor in this process. An index functions like a smaller table in ordered sequence (Paller, 1997) and provides direct access to the rows of interest to a user without having to scan all the rows of the entire table. Unlike ODS databases, DW databases use indexes extensively. To ensure an optimal indexing methodology, multiple indexes are created on most of the dimension tables. Livingston and Rumsby (1997: 191) said "*To ensure fast access to this high volume of data, a good strategy would be to put indexes on every single column of each dimension table*". The star schema structure is based on the direct relationships between the fact and the dimension tables. An important aspect of this structure is the use of multiple table-joins and indexing. The join operation is accelerated by using indexes.

[3] Adapted from (Sorensen and Alnor, 1999).

Although (Livingston and Rumsby, 1997) *said that the most relevant indexing approach for the DW is to index all columns in the dimension tables and all foreign keys in the fact table,* as this will improve the performance of the DW by reducing query processing time, Kimball (1996) recommended the use of indexing on some columns in the dimension tables which are extensively used by queries. Kimball said that there is no need for indexing columns that are of little use (i.e. first name, zip code, ...etc) because of the cost of indexing (e.g. the increased number of I/O operations). Rather than indexing all the fact table columns, the index relies heavily on the primary key foreign key match. When the query is processed the primary keys of all relevant dimension tables are concatenated, and then the matching rows in the fact table are found without having to scan all of the fields in the fact table. Using this primary key foreign key index technique improves the performance of the queries taking place in the DW database (Berson and Smith, 1997; Livingston and Rumsby, 1997).

3-9-9 Star Schema example

The star schema described in figure (3-6) consists of four tables. The *fact* table is the **Student_record** whilst there are three *dimension* tables: **Course**, **Student_data**, and **TIME**. Where time is always part of each DW. Each dimension has a primary key **(PK)**: Course (course key), Student_data (student key), and TIME (time key). The PK of the fact table is obtained by the concatenation of all the dimension tables PK's (course key + student key + time key).

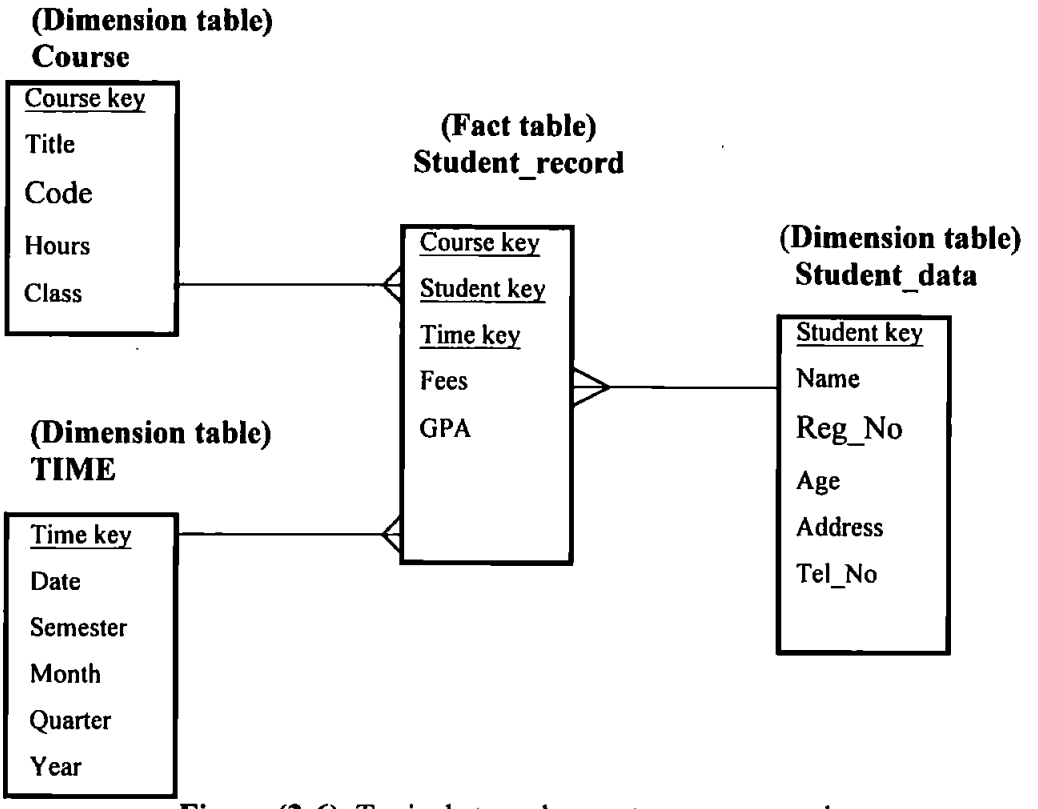


Figure (3-6). Typical star schema structure example.

3-9-10 Summary of the Star Schema Structure main characteristics

Based upon the DW literature (Wixom and Watson, 2001; Humphries, et. al, 1999; Sørensen and Alnor, 1999; Hadden, 1998a; Hadden, 1998b; Firestone, 1998; Adamson and Venerable, 1998; Mimno, 1998; Devlin, 1997; Livingston and Rumsby, 1997; Kimball, 1996) the Star Schema Structure has the following characteristics:

1. the Star Schema consists of both FACT and DIMENSION tables;
2. the FACT table has relationships with the DIMENSION tables. However, there are no relationships between the DIMENSION tables;
3. foreign keys relating a row in the FACT table to a DIMENSION table can not be NULL;
4. optional relationships causes FACT table problems called FAKE rows;
5. the FACT table may have multiple foreign keys to the same DIMENSION table;
6. the FACT tables are SPARSE. SPARSITY means that there is no row in the fact table for each possible combination of the DIMENSION tables;

7. the FACT table is DEEP i.e. consists of thousands or even millions of records.
8. all records in the FACT table are numeric;
9. the most useful FACT table should be additive;
10. in some star schema designs it is not necessary to store any attributes in the FACT table other than the foreign keys. In this case it is referred to as FACTLESS FACT table;
11. the DIMENSION tables are WIDE (consists of tens or hundreds of attributes) either numeric or text;
12. the DIMENSION tables are always DENORMALIZED;
13. the primary key of the FACT is a concatenated/composite key and consists of all DIMENSION primary keys;
14. the primary key of the dimension table may be different from its key in the operational systems;
15. there is often a TIME DIMENSION in most DW;
16. the GRAIN of the FACT table is the level of detail in the DW;
17. the best INDEXING approach for the DW is to index both the FACT table foreign keys and all the DIMENSION table columns;
18. aggregation and grouping of data is achieved by using SQL.

3-10 Data Warehouse components

A DW consists of six components, these components are; data sources, data extraction and transformation tools, data modelling tools, central repository, target DB, and front-end tools. The following sections provide the details of these components.

3-10-1 Data source

These are the ODS' databases, external, personal or archival data sources. Different data source formats can be used as sources of data, for example VSAM, IMS, RMS, DB2, relational, flat files and other formats (Mimno, 1997);

3-10-2 Data extraction and transformation tools

These are used to extract data from the data source files, clean and transform the data, and to ensure that all the relevant data required by the users is available in the DW.

The extraction can be done using a standard RDBMS (e.g. ORACLE, SYBASE, INFORMIX, DB2, SQL, ACCESS).

The data transformations might include aggregating, inserting default values, sampling or summarizing the data to reduce the size of the DW (Widom, 1995). Various tools (e.g. Extract, Integrity Data Reengineering, Platinum Warehouse) have been developed to support the transformation process.

During the transformation process the operational system fields/attributes are copied to the DW. Many transformation types are available to perform this mapping; field splitting, field consolidation, standardization, and deduplication. The following table (3-3) provides examples.

Transformation type	Operational System	DW/Data mart
Splitting	Address	Street No City Country
Consolidation	Street No City Country	Address
Standardization	Address	Address
Deduplication	**ODS 1: Booking DB** Student Name **ODS 2: Fees DB** Student Name	Student Name

Table (3-3). Field/attribute ODS to DW mapping[4].

[4] Adapted from (Humphries, et al., 1999).

When the data collection and extraction process starts some problems may appear. Table (3-4) contains possible sources of problems and potential solutions.

The problem	Reason	Solution
Incorrectness	-Inaccuracy in data entry.	-Systematic way to ensure data quality, double check, hash totals, coding, or check digits.
Lack of timeliness	-Data generation methods are not fast enough.	-Generate the data on predefined times.
Improper measurement	-Data collection not consistent with the purpose of the DW.	-Change the measurement methods.
Lack of required data	-Required data in not stored or no one stored it before.	-Estimate the data values if relevant, or start the store procedure now.

Table (3-4). Data problems[5].

3-10-3 Data modelling tools

These tools are used to prepare the DW structure from both the data source and the target data warehouse database;

3-10-4 Central repository

This component is used to store the metadata (data about data). Metadata describes the transformation between the source and the target databases.

3-10-5 Target DB

This is the DW database, where the data of interest will be stored. The target database model can be a conventional relational database, *proprietary, or multi-dimensional* (Mimno, 1997). In many cases organisations use the standard RDBMS to build the DW target DB.

[5]Adapted from (Turban and Aronson, 1998).

Proprietary DB supports high performance tools including file segmentation and partitioning, stored procedures, custom data warehouse functions, support of iterative queries, and high performance query processing. As with all data warehouses the performance decreases as the size increases.

Multi-dimensional DBMS look at the tables from more than two-dimensions called *Cubes*. These are optimized for on-line analytical processing (OLAP) (Adamson and Venerable, 1998). They are implemented through the use of table groupings, nested tables, list-derived functions and advanced indexing options. The software that follows this approach is called a multi-dimensional database management system (MDDBMS). MDDBMS allows the user to look at the details of the business in terms of measurements which quantify what the business is doing. The lowest level of detail contains no aggregation. Typically, data is stored at the detail level are then aggregated by day, week, month, or quarter. Measurements may be as many as 20 dimensions, however users frequently work with two to four dimensions (Edelstein, 1997). Figure (3-7) provides an example;

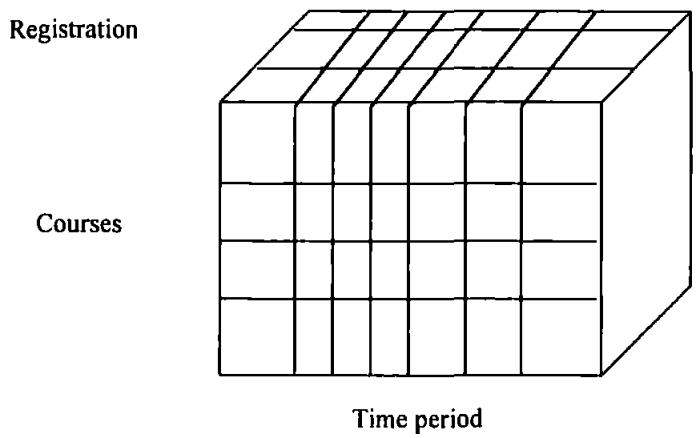


Figure (3-7). Three-dimensional view.

Different MDDBMS engines have different capabilities in terms of the number of fact and dimension tables they can store. Most of the MDDBMS have a DB size capacity below 100 gigabyte (Humphries, 1999). Examples for these products are Essbase (Arbor), Enterprise (Brio), and Express Server (Oracle).

The MDDBMS allow the users to change their perspective on the data interactively at a high speed, adding and removing attributes, and receiving results concurrently. However,

they do not perform well when there is a large number and size of records. This may lead to a cube with slow performance and high space consumption. These problems can be resolved when implementing the DW using a RDBMS, which have been enhanced to support the specialized requirements of DW. Examples of these systems are MS-SQL Server 7.0 (Microsoft), Sybase SQL Server (Sybase), Informix (Informix), Oracle 7 & 8 (Oracle), and IBM DB2 (IBM). Users can also retrieve data in multi-dimensional views using any front end access tool from the RDBMS that stores the DW. This gives another point of strength to the use of the RDBMS. This approach is generally used when an organisation needs to perform some deep analysis on a subset of a very large data set (Adamson and Venerable, 1998; Mimno, 1997).

3-10-6 Front end

These are the tools used to analyze the data stored in the DW database. These tools include:

- General-purpose relational data access;
- Data mining tools;
- DSS;
- EIS;
- Web tools that perform search and query in the WWW environment.

For some applications a mix of the front-end tools may be required (Laudon and Laudon, 2000; Mimno, 1999; Mattison, 1997; Berson and Smith, 1997; Mimno, 1997; Adriaans and Zantinge, 1996).

3-11 Client/Server structures for supporting data warehousing

Client/Server architecture consists of a number of workstations (*Clients*), one or more higher configuration workstation(s) (*Server(s)*), and a local area network (LAN) connecting them all together (Edwards, 1999; Orfali, 1999; Delis and Roussopoulos, 1992). Client/server applications involve the dispersing of the software over several

computers and creating a seamless environment for the end-users so that it appears as though they are working on just one system.

Client/Server architecture has changed the way DBMS are build and operate. The server hosts the DBMS and the DB(s), whilst the client runs the application locally and sends requests to the server in order to retrieve data. The server replies to the client's request and then the client is able to perform the required analysis and reporting functions locally.

The processing of the graphical user interfaces or other visual techniques can be carried out on this local machine whereas a specific database server handles all the database tasks. In this way the database server can be optimized to perform these tasks. There are three basic classes of application databases that can be facilitated by the Client/Server architecture:

1. Read Transaction (RT) databases with a lot of read-only clients connected to the server. This class includes Internet ftp sites, libraries, CompuServe and data warehouses (Orfali, 1999; Berson and Smith, 1997; Delis and Roussopoulos, 1992).
2. Constant number of update transaction (CU) databases with many read-only users and a constant number of updates. This class includes DBs such as those storing stock market prices where only a few users have the right to update while the others are performing read-only transactions.
3. Variable number of update transaction (VU) databases in which both number of reads and updates are proportional to the total number of clients. This class includes TPS.

The end-user can work on the local workstation or connect to a display server that has access to the data warehouse running on one or more database servers. Data can also be extracted to a local database system and then processed. This is all possible within a client/server environment because each computer is set up to fully optimize the end-user's application.

Moreover, the data warehouse is often a client-server application. *Of all the techniques currently available, client/server represents the best choice for building a data warehouse*

(Orfali, et al., 1999; Edelstein, 1997; Delis and Roussopoulos, 1992). The role of the client depends on the DW architecture.

3-11-1 The DW architecture

Within a client/server data warehouse architecture there are two alternative structures; two-tier and multi-tier.

3-11-1-1 The two-tier DW

The two-tier (*2-tier*) DW architecture or as sometimes called *the fat client* model (Edwards, 1999; Edelstein, 1997), in which clients' functions include GUI presentation logic, query definition, data analysis, report formatting, summarization, and data access, whilst the DW server performs data logic, data services, metadata maintenance and the file services. However, the two-tiered architecture lacks scalability and flexibility (Edwards, 1999). As the number of users increases the data access requirements imposes heavy burden on the server and the performance degrades (Mimno, 1997). Source data and the data warehouse DB reside on the server (tier 2), whilst business rules that are shared across the organisation and graphically oriented end-users run on LAN-based workstations (tier 1).

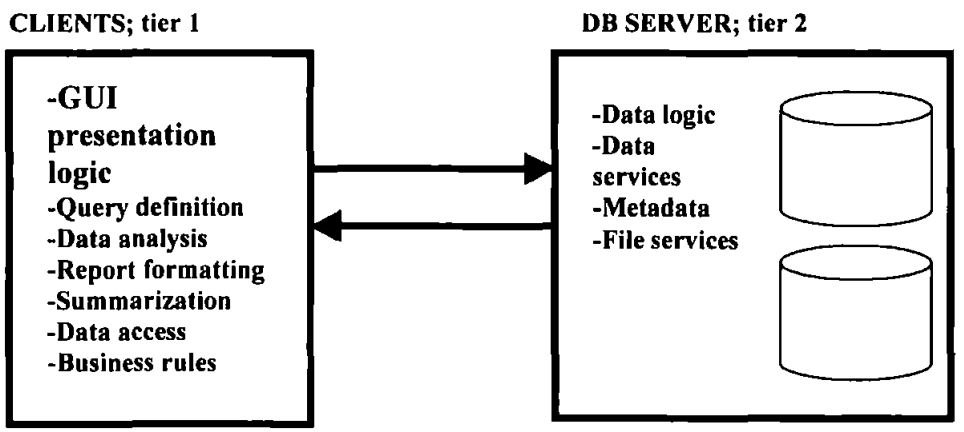


Figure (3-8). Two-tier data warehouse architecture.

3-11-1-2 The multi-tier DW

The multi-tier (*3-tier*) DW architecture or as sometimes called *the thin client* model, handles the scalability and flexibility problems through the application servers. Application servers perform data filtering, summarization, aggregation, support metadata, data access,

and provides multi-dimensional views. The multi-tier architecture reflects the multi-tier client/server model (Edwards, 1999; Edelstein, 1997; Dewire, 1998). Source data resides on the server (tier 3), data warehouse DB and business rules that are shared across the organisation are stored in a DB Server (tier 2), and graphically oriented end-users run on LAN-based clients (tier 1).

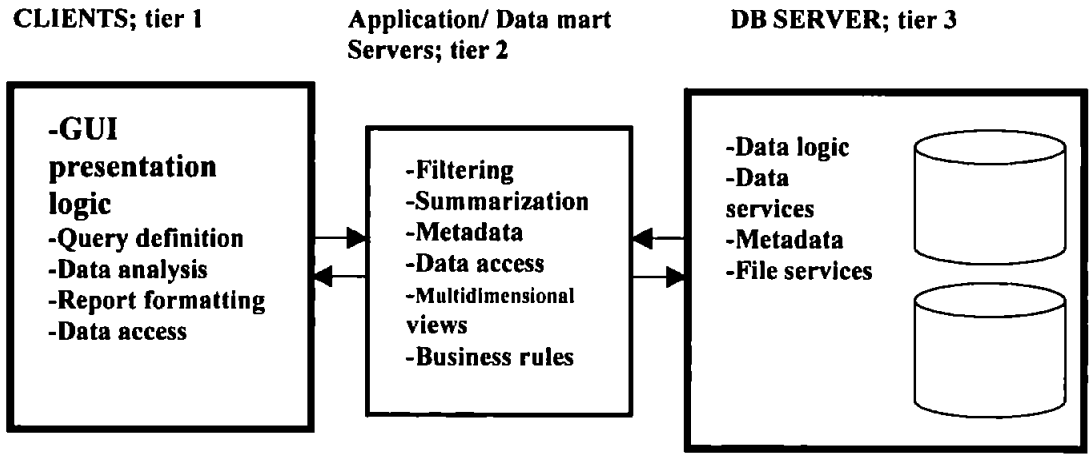


Figure (3-9). Multi-tier data warehouse architecture[6].

3-11-2 Comparison between the DW architectures

> *"When a client/server was a departmental or campus-based phenomenon, the shortcomings of 2-tier were not very important. They certainly did not outweigh the advantages provided by 2-tier's ease of development. But as client/server grew up to run mission critical applications-especially those of intergalactic proportions- 3-tier became essential"* Edwards, 1999: 9

The following table (3-5) includes a comparison between the two architectures.

[6] Adapted from (Edelstein, 1997; Livingston and Rumbsy, 1997).

Criteria	Two-tier (2-tier)	Multi-tier (3-tier)
System administration	*Complex* More logic on the client to manage	*Less complex* Can centrally be managed on the server
Security	*Low* Data level security	*High* Service or method level
Performance	*Poor* Many SQL statements are sent over the network	*Good* Only service requests and responses are sent over the network
Ease of development	*High*	*Low* Tools are emerging to enhance this process
Legacy application integration	*No*	*Yes* Via gateways
Heterogeneous DB support	*No*	*Yes*
Internet support	*Poor* Internet bandwidth limitations make it harder to download fat clients	*Excellent* Thin clients are easy to download as applets or beans

Table (3-5). Two-tier Versus Multi-tier architecture[7].

The comparison shows that the multi-tier architecture has advantages over the two-tier, and this justifies why the multi-tier architecture is most frequently used in practice.

3-12 Data warehouse development approaches

Two basic approaches are used to build a data warehouse, known as the *'top down'* and the *'bottom up'* approaches (Edelstein, 1997; Berson, 1996).

[7] Adopted from (Edwards, 1999: 9).

1. *The 'top down' approach.* The organisation has developed an enterprise data model, collected enterprise-wide business requirements. It then builds an enterprise data warehouse with subset data marts.
2. *The 'bottom up' approach.* This implies that the business priorities result in developing individual data marts, which are then integrated into the enterprise data warehouse.

Given the way in which computer systems have been developed, the bottom up approach is probably the more realistic, but the complexity of the integration may become a serious obstacle, and the warehouse designers need to carefully analyze each data mart for integration purpose. If the set up of a datamart is to be used with data mining techniques, then optimizing the local databases is important. The designers must ensure that the hardware and database requirements that have been established are adequate for this purpose.

3-13 Users of the Data Warehouse

Taking into consideration the nature and characteristics of the DW and its strategic use and the historic data it contains reveals that the typical users are:

- Those who need certain amount of data in a special format for the reason of summarizing, and aggregating data;
- Those who deal with analyzing and displaying historical data;
- Those who need to reply to the frequently asked queries-FAQ;
- Those who need continuous accessing of a certain data and exception reporting (Taha, et al., 1997).

This list reveals the fact that *managers* and *executives* (decision makers in organisations) are the primary candidates to use the DW because they need to monitor and track what is happening in their internal and external environments. These decision makers also use the

DW to predict the future performance of the company in order to develop correct strategic plans and implementation processes. This prediction process uses the historical data, which is available from the DW (Berson and Smith, 1997; Adriaans and Zantinge, 1996).

3-14 DW size and number of users

A survey (Teklitz, et at., 1999) investigated more than 3000 DW users. The study had two dimensions; size and number of users. Figures (3-10 and 3-11) depict the study results.

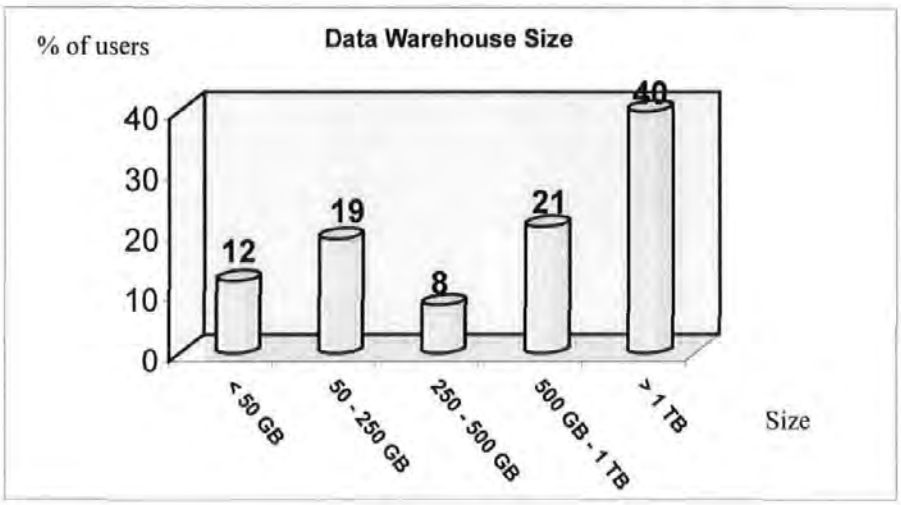


Figure (3-10). DW Statistics.

Figure (3-10) shows that a majority of 61% (21 + 40) of the DW users have a DW size ranging from 500 GB up to > 1 TB.

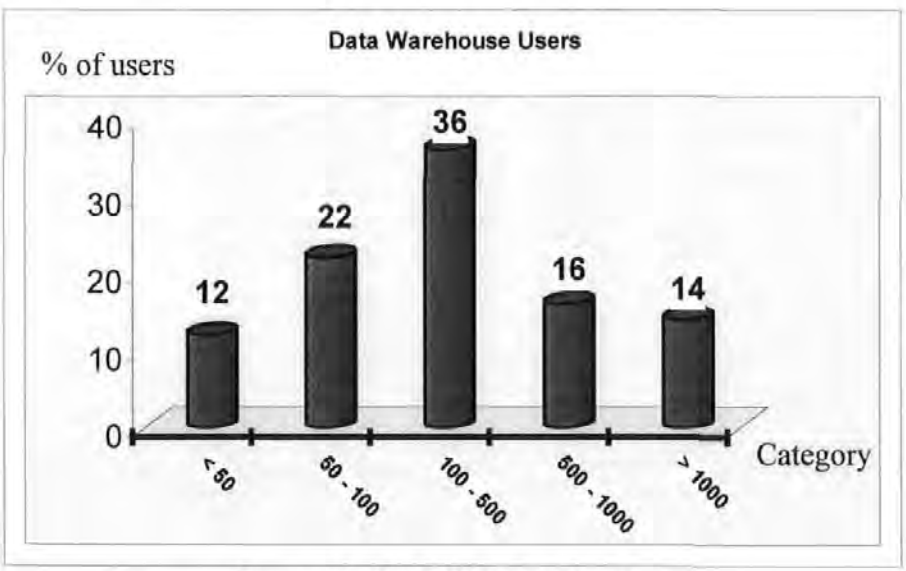


Figure (3-11). DW Statistics-1.

Figure (3-11) shows that a majority of 66% (36 + 16 + 14) of the DW systems have over 100 users.

Teklitz, et al. study results emphazised that there is a trend by which the number of DW users and DW size are growing in organisations.

3-15 The data warehouse development strategy

According to (Humphries, et al., 1999; Hadden, 1998b; Levin, 1997), a DW development strategy is required at the early stages of the project; this strategy is described in a *DW development strategy document*. A DW development strategy document contains the following items:

1. *The project objectives and benefits*. The business objectives (i.e. better business understanding, higher decision quality, enhancing customer satisfaction..etc) of the DW project should be clearly identified and related to the organisation objectives. Also the expected DW project benefits (i.e. more cost-effective decision making, supporting the corporate strategy, reduces redundant processing, improving productivity..etc) need to be stated;
2. *The duration of the DW development project*. On average a DW project takes three to six months to be completed (Levin, 1997). However, for large organisations that intend to build EDW directly, the duration of the DW project could be years depending of the variety of the data sources, heterogeneity of the hardware and software platforms, and the application complexity;
3. *The approach used*. That is, either the top-down or the bottom-up approach;
4. *The DW development team*. Humphries, et al. (1999: 77) said, "*Every data warehouse project has a team of people with diverse skills and roles*". According to Humphries, et al., the DW team consists of *ten roles*; *five* of which should be fulfilled by internal staff (i.e. Steering committee, User reference group, Warehouse driver, Source system

DBA, and Project sponsor), whilst the other *five* roles could be fulfilled by either internal or external staff (i.e. Warehouse project manager, Business analysts, Warehouse data architect, Metadata administrator, and Warehouse DBA);

5. *Data warehouse rollouts' plan.* A DW project can not meet the users' requirements in one step, hence, prioritizing the various users' requirements and assigning them different *DW rollouts* is a realistic approach. DW rollouts enable organisations to divide the DW project into successive phased tasks. Applying this phased DW rollouts enable organisations to increase the functionality of the DW and to lower the overall DW project risk (Humphries, et al., 1999; Hadden, 1998b);
6. *Users.* Users of the DW are mentioned because based on their information needs and requirements the DW will be built;
7. *Define the data warehouse architecture.* The data warehouse architecture includes defining: which DBMS (e.g. MDDBMS or RDBMS) will be used, the hardware specifications, the client/server architecture, the report generation tool, front-end tools...etc;
8. *Identifying data sources.* The ODS that represent the data sources for the DW should be identified. Also if external and/or personal data sources are to be used they need clear identification;
9. *The DW updating policy.* Data warehouses collect and store data for querying and reporting purposes, after some time, the data may no longer be needed and/or may not be of interest to the DW users, when this situation occurs the DW needs to be updated (Garcia-Molina, et al., 1998). Updating makes the DW up-to-date by copying data from the ODS to the DW. Widom (1995) said that in many cases it is not desirable to keep data "*forever*", techniques and policies are required to ensure that outdated data is automatically and efficiently purged from the DW. Updating the data warehouse depends on the following factors:

a. *The application type*. The type of application drives the updating strategy of the DW. That is, in a *Stock market DW* the securities' data needs to be updated daily or weekly, however, in a *Super Market DW* the products' data could be updated weekly or monthly, and in a *University DW* the students' data could be updated every semester or academic year;

b. *The users' needs*. Different users have different needs which the DW updating strategy should respond to. For example, in a University DW Deans are interested in the Admission Curve report which is generated every academic year, whilst, Registrars are interested in the Course Bookings Report which is generated every semester. In this situation if Deans and Registrars use the same DW, it would be updated every semester. However, if they use different data marts the Deans' data mart could be updated every year, whilst the Registrars' data mart could be updated every semester.

3-16 DW development guidelines

1. *Data and Database heterogeneity*. Data heterogeneity is the difference in how the data is defined in different database models. For example, the different ways of modelling the same fact, different attributes for the same entity. Because the data warehouse may also contain the organisation's historical data, the DW must be capable of storing and managing large volumes of data which may be in different models. Database heterogeneity appears when the DW deals with different DBMS models and/or vendors (Srivastava and Chen, 1999);
2. *Outdated data*. It is undesirable to keep data forever. A strategy must be set to ensure that outdated data is efficiently purged from the data warehouse (Widom, 1995);
3. *Start with the correct sponsorship chain*. One of the most important factors in the DW success is the sponsorship. The right sponsorship chain includes the executive sponsor

who is a key person above the DW manager. They should be aware of the technology and have the authorization to assign the required facilities and funds;

4. *Set achievable objectives.* The development of the DW includes the planning and the implementation phases. In the planning phase the objectives of the DW are set and defined for users. In the implementation phase these objectives should be achieved. However, statistics show that in many occasions there is a gap between what was planned to be achieved and what was actually implemented (Barquin, et al., 1997);
5. *Data selection.* Data that is to be stored in the data warehouse should be selected carefully because only relevant operational database fields should be copied to the data warehouse. Irrelevant data increases the size of the database, which will passively affect the performance of the DW and definitely drains the hardware resources (Barquin, et al., 1997). In some situations the DW will hold historical data for up to 20 years, and during this period the size of the DW will increase and this may adversely affect its performance;
6. *Continuos maintenance.* Once the DW project is completed the development team should consider the future maintenance of the DW. Maintenance includes the performance of the DW, the data items to be added or deleted, requirements of new users, and the response time problems;
7. *Focusing on both internal and external data.* The primary users of the DW are the users of DSS and EIS, because these are the applications that use DW for strategic data usage. Executives and top managers are taking long-term decisions that affect the overall future of the organisation, in doing so they use both internal and external data sources. External like competitors, customers, demographic data, or government data sources. When they are using the DW, through their DSS/EIS application that is enhanced with DW component, they should be able to retrieve either internal or external data;
8. *Choosing the DW manager.* The DW manager should have skills in two areas, the first is the technology of data warehousing and the second is the communication skills. The

communication will frequently happen between the DW manager and the users of the system to respond to their requirements;

9. *Data Warehouse administration.* According to a study by SAS Institute (1999), administering the DW environment is a job which is critical to the success of the DW project. In an ideal DW environment, the administration team is responsible for tuning up the following components together: ODS, DW DB, users, backup/recovery, DBMS, data exploitation tools, data transformation tools, security, performance, CASE tools, and the update policy.

Chapter summary

- This chapter has focused on the need for a DW to meet the increasing data needs of organisations which could not be satisfied with the TPS.
- A DW extracts data from a variety of databases both internal and external to the organisation.
- The DW stores this data so that it can be effectively analyzed using EIS and DSS.
- The purpose of the DW is to establish data repository that prepares the operational database in organisations in an accessible and ready-to-use format.
- There are four characteristics that generally describe a DW. A DW is time-variant, non-volatile, subject-oriented and integrated.
- DW benefits include leveraging the decision maker's productivity by providing accessible data in a ready to use format and isolating the operational databases from the query processing that keeps the performance of the TPS up.
- An EDW contains large number of data fields and millions of data records about the entire organisation.
- Smaller, local data warehouses that are classified by subjects are called data marts.
- The three fundamental database models are the relational model, hierarchical, and the network. There are also new database models these are the object-oriented, multi-media, and the star schema structure. DW database can be a relational model or star schema structure. The star schema structure is the best for DW design.
- DW components are the data source, data extraction and transformation techniques, data modelling tools, central repository, target DB, and front-end tools.
- The data warehouse application is a client-server application, preferably multi-tiered because of the following advantages; highly secured, less complex, has a high

performance, and is able to support heterogeneous DB models in addition to the Internet applications' support.

- The basic approaches used to build a data warehouse are the top down and the bottom up, the bottom up is easier for organisations.
- A DW development strategy document should be exist at the early stages of the DW project.
- Another useful technique used for extracting information and knowledge from the DW is called data mining. It is the process of finding hidden knowledge and unknown facts and trends in data. Data mining techniques are part of the knowledge discovery in database (KDD) process. KDD and data mining will be covered in the following chapter, revisiting its relationship with DW.

Chapter four

KDD

Techniques

This chapter will cover the knowledge discovery in database (KDD) process. It begins with evaluating the different types of knowledge and how to reach them using various types of techniques. Then KDD process and its importance will be identified. Different terminology for the knowledge discovery process will be discussed with particular emphasis on data mining. The distinction between KDD and data mining will be clarified by showing the place of the data mining in the KDD process. The tasks and goals of data mining are evaluated. Data mining techniques including query tools, visualization, on line analytical processing (OLAP), association rules, decision trees and rules, artificial neural networks (ANN), clustering, genetic algorithms and probabilistic graphical dependency technique will be discussed and evaluated. The role of the user in the KDD process will be identified. To place the entire KDD process in context, it is applied to a sample data set drawn from the records of the Arab Academy for Science & Technology and Maritime Transport (AASTMT). Finally relevant research and application challenges facing KDD will be identified and discussed.

4-1 Knowledge types

There are four broad categories of knowledge within organisations.

4-1-1 Shallow knowledge

This refers to the representation of surface level knowledge. Shallow knowledge describes the input/output relationship of the system (Turban, 1993). The information retrieved from this source of knowledge is simple and can be easily obtained by using normal search routines e.g. Structured Query Language (SQL). Shallow knowledge is inadequate for explaining complex situations or where explanation is part of the solution.

4-1-2 Multi-dimensional knowledge

This type of knowledge can only be obtained using specific techniques such as on-line analytical processing (OLAP) and visualization. OLAP tools offer the ability to explore rapidly all sorts of clustering and orderings within the data. The major benefit gained

behind the use of OLAP is that it is designed for multi-dimensional searching. Further, the OLAP technique is used to make the output multi-dimensional. It is also important to realize that most of what could be done with OLAP tools could also be carried out using SQL. However, SQL takes longer and deals with a fewer number of tables than using OLAP. Visualization techniques may be used throughout the multi-dimensional knowledge exploration process (Chen and Paul, 2001).

4-1-3 Hidden knowledge

This knowledge that can be found by using groups of the data mining techniques, such as Artificial Neural Networks (ANN), OLAP, Genetic Algorithms (GA), and decision trees. Combining more than one data mining technique for the same data set should take into consideration the match between the type of data and the characteristics of the technique. We also can not combine all of the data mining techniques together due to the differences in their characteristics that may sometimes give contradicting results (Burn-Thornton, et al., 1998; Fayyad, et al., 1996).

4-1-4 Deep knowledge

Deep knowledge refers to the internal and causal structure of a system, and the interaction between its components relative to certain business applications. The entire KDD process is able to handle this type of knowledge. Where there is a DW installed and DSS front-end tools exist to handle this knowledge output. The four types of knowledge are summarized in figure (4-1).

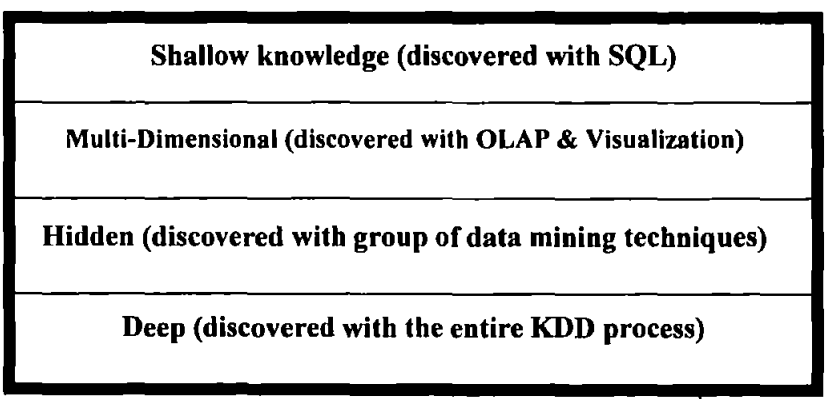


Figure (4-1). The Different types of knowledge[1].

[1] Adapted from (Adriaans and Zantinge, 1996).

In the next sections, the definition and details of the knowledge discovery process will be introduced then discussion of the various data mining techniques, followed by an example that uses different data mining techniques to find different types of knowledge.

4-2 The emergence and definition of the KDD process

The term knowledge discovery in database (KDD) was coined in 1989 to point to the process of finding knowledge in data (Fayyad, et al., 1996). KDD is also defined as the process of finding patterns of hidden information or unknown facts in the database (Riedel, et al., 2000). Traditionally the finding of useful unknown patterns and hidden information in raw data has been given many titles including knowledge discovery in database, data mining, data archaeology, information discovery, knowledge discovery or extraction, and information harvesting (Adriaans and Zantinge, 1996). There are two reasons for this lack of consensus; the first is the novelty of the KDD and the second is the multi-disciplinary nature of KDD. Multi-disciplinary means that KDD has been developed in many disciplines e.g. statistics, geography, medicine, and computing (machine learning, artificial intelligence (AI), databases, data warehousing, expert systems, knowledge acquisition and data visualization) (Fayyad, et al., 1996). Although, the field of KDD was founded in many disciplines, it is gaining its character on its own and now stands by itself (Ramakrishnan and Grama,1999).

The interest in KDD has increased and this is demonstrated by the increasing number of forums and workshops (Fayyad, et al., 1996). Another sources of interest are the various publications and special issues that document some of the KDD features and foundations (Aas, et al., 1999; Piatetsky-Sharipo 1995; Cercone and Tsuchiya 1993; Parsaye and Chignell 1993; Piatetsky-Sharipo 1992; Inmon and Osterfelt 1991). Interest in the KDD process has widened with the Web's emergence as a large distributed online data store or repository and the realization that this can de used for extensive commercial purposes.

4-3 KDD or data mining

Scientists have used the terms KDD and data mining interchangeably (Ganti, et al., 1999). However, others have stated that data mining is a step in the KDD process (Adriaans and Zantinge, 1996).

For the purpose of this research data mining is considered a step in the KDD process. This is because the main objective of this research is to build DSS using the KDD process, not to develop a new data mining technique. Thus, the research deals with KDD as the overall process of discovering useful knowledge from data whilst data mining points to the application algorithm or technique that is used for extracting patters and unknown information from the raw data. So, the KDD process will get knowledge or information from the data mining techniques applied to a certain application.

4-4 The KDD process

The KDD process is interactive, iterative, and involves a great deal of user-interference. Brachman and Anand in 1996, defined the practical view of the KDD process as follows:

1. *Developing an understanding of the application domain.* This is an important step because it determines the goals of the KDD application. Based on these goals relevant data mining techniques can be employed. However, there is no single technique that best fits all application domains. In addition the overall performance of the KDD process will be evaluated based on the domain and its goal(s);
2. *Creating a target data set.* Selecting the data set, or focusing on a subset of the database on which discovery will take place;
3. *Data cleaning and preprocessing.* This includes basic operations such as the removal of noise (if relevant), and deciding on the strategy for dealing with missing data items. Example of the strategies that might be used here are neglecting the incomplete data records or setting missing values to null. Since both affect the accuracy of the output

knowledge this decision is important and the strategy used is often based on the volume of the incomplete data and its importance;

4. *Data reduction and projection.* This involves finding useful features to represent the data set depending on the stated goals. For example if the goal of the KDD task is to determine and predict the students' academic performance, not all of the student's record is important. For example, the student's address, telephone number or height and weight are irrelevant ;
5. *Choosing the data mining goal and task.* Choosing the data mining goals and tasks are related to the KDD goals that need to be achieved. The data mining goals are either prediction, description, or both. However, the data mining tasks are classification, clustering, summarization etc. Choosing the data mining goals and tasks is related to the goals of the KDD process;
6. *Choosing the data mining technique(s) or algorithm(s).* This is about selecting the methods to be used for searching in the data for patterns and hidden information. This includes deciding on the relevant models and parameters. The chosen techniques should satisfy the data mining tasks and goals, and the KDD goals;
7. *Data mining.* Searching for patterns in the data sets using analysis methods and models such as regression, clustering, SQL, visualization, decision trees and others;
8. *Interpreting the information gained by the mining techniques.* The output of the mining techniques should be evaluated and presented so that they are understandable and consistent. In practice, iterations from steps 1 to 7 often occur when applying KDD;
9. *Consolidating the discovered knowledge.* Reporting the knowledge to the interested parties and checking the discovered knowledge with the previously known knowledge.

During the KDD process, particularly the data mining step, it is necessary to search the database(s) of the organization. Either the enterprise data warehouse, or the data mart of a department could be used to enhance the output of the KDD process by providing a wealth

of historical information to the mining techniques (Berson and Smith, 1997; Adriaans and Zanting, 1996). The following figure (4-2) illustrates the whole KDD process.

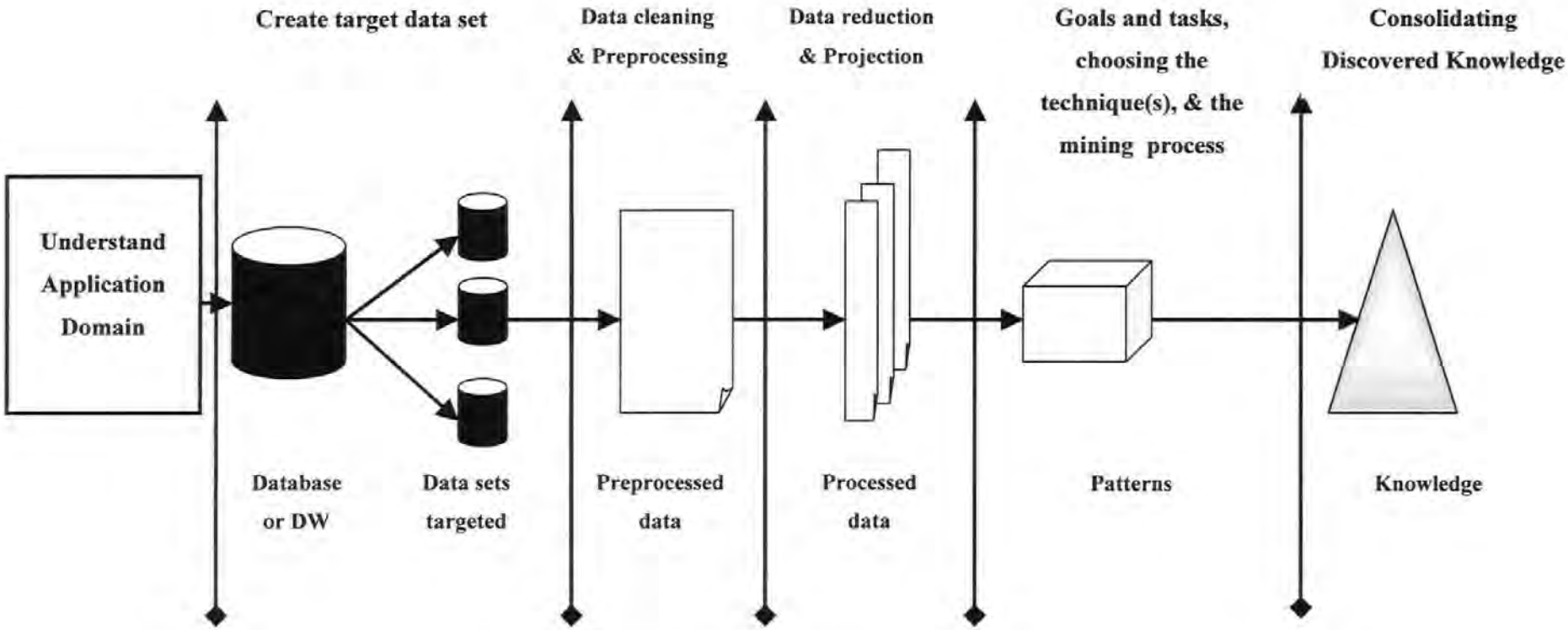


Figure (4-2). The KDD process overview[1].

[1] Adapted from (Fayyad, et al., 1996).

From the work of Brachman and Anand summarized in figure (4-2) we can deduce that:

1. KDD is an entire process that should be applied from the application domain identification step until the evaluation of the discovered knowledge;
2. The discovered knowledge should be then utilized in a suitable front-end tool like EIS or DSS;
3. A DW will enhance the KDD results, furthermore, a DW without a front-end like EIS or DSS to assist end-users in the decision making process is probably an unsuccessful warehouse (Barquin, 1997; Paller, 1997; Taha, et al., 1997);
4. As a result of this work, organisations would get the maximum benefits from their business applications if they include these components (i.e. DSS, DW, and KDD) together.

4-5 The primary tasks of data mining

In practice there are two high-level but fundamental goals of data mining *prediction* and *description* (Fayyad, et al., 1996). Prediction is the use of some variables or data fields in the database to predict the unknown future values of the other variables or data fields of interest. Description focuses on finding understandable patterns describing the data set. The relative importance of these goals varies from one application to another, for some applications prediction is more important than description and for another applications description is more important. In the KDD process, description is much more important than prediction (Fayyad, et al., 1996), whilst in machine learning and pattern recognition applications prediction is the principal goal. Both goals, prediction and description are achieved using the following data mining tasks.

1. *Classification.* This is a learning function that classifies data records into one of several predefined classes. A practical example of classification would be to use students' grade point average (GPA) to determine whether or not they had a scholarship.

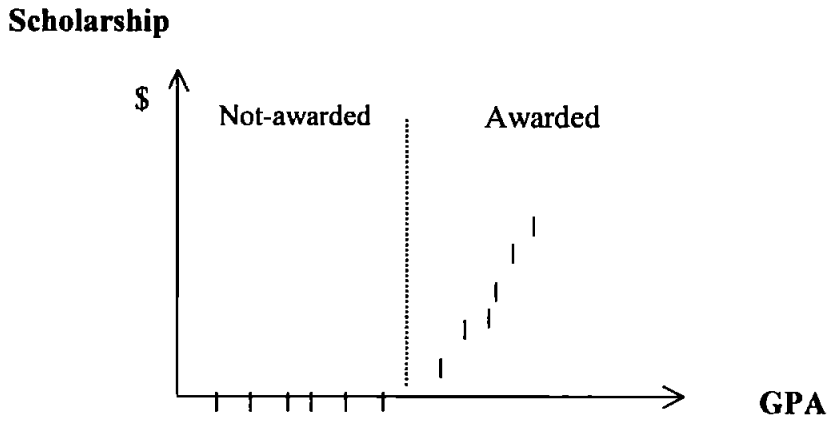


Figure (4-3). Simple linear classification boundary for artificial students' data sets.

Figure (4-3) shows the simple partitioning of students into these with scholarships and those without scholarships based on their GPA. In many real life applications the line that best separates the classes will not be linear. The university could use this line to forecast the number of scholarships they will give;

2. *Regression.* A regression model is a mathematical equation that provides predictions of the values of one variable (dependent) based on the known values of one or more other variables (independent or predictors). If one predictor is considered then the regression is called simple linear regression, if more than one variable is used then the regression is called multiple linear regression (Canavos and Miller, 1995). Regression models have been used extensively in many disciplines including finance, computing, engineering, medicine etc. Regression can be used to predict the number of new students that will join the university, predict the census of a certain country, and loan predictions of a certain bank.

In figure (4-4) the regression line establishes the relationship between the value of a scholarship and the students' GPA. The fit of the regression line is not perfect due to other variables that could have an effect on the scholarships awarded;

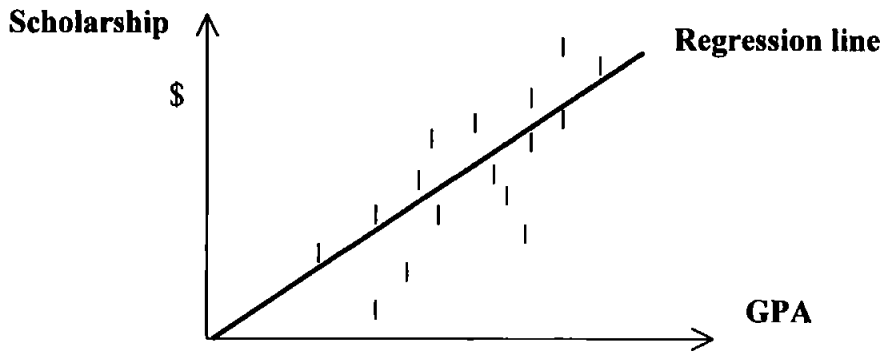


Figure (4-4). Simple linear regression for artificial students' data sets.

3. *Clustering*. Clustering, Q-analysis, typology, grouping, clumping, numerical taxonomy and unsupervised pattern recognition are used interchangeably to refer to the same thing. Clustering is the process of producing classifications from initially unclassified data (Larsen and Aone, 2000; Guha, et. al, 2000; Everitt, 1980; Anderberg, 1973). It is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. The clusters may be mutually exclusive or exhaustive or may be overlapping (Fayyad et al., 1996). Where clusters overlap, data points lying in these regions belong to more than one cluster;

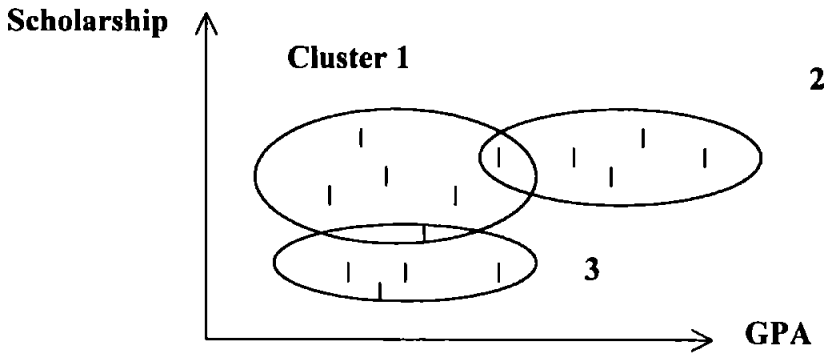


Figure (4-5). Simple clustering (3 clusters) for artificial students' data sets.

4. *Summarization*. It is the process of finding a compact description for a subset of data. For example the mean and standard deviation of all data fields of interest, the discovery of the relationships between variables, and the use of multivariate visualization

techniques. Summarization is often applied to interactive exploratory data analysis and report generation;

5. *Dependency modelling*. This consists of finding a model that describes significant dependencies between variables. Dependency models fall into two categories:
    a. *structural*. Determines which variables are dependent on each other;
    b. *quantitative*. Shows the strength of the dependencies using a numerical scale e.g. *probabilistic dependency networks* which are used in medical expert systems, modelling of the human genome, and information retrieval;
6. *Change and deviation detection*. This focuses on discovering the most significant changes in the data from previously measured or normative values.

4-6 The data mining algorithm(s)

Once we have defined the principal task(s) and goal(s) of the data mining process, the next step is to develop the algorithm(s) that will achieve both the task(s) and goal(s). Cormen, et al. (2000: 1) said "an *algorithm* is a well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output". In other words, an algorithm is a series of steps that instruct the computer how to resolve a certain problem by transforming the input into output.

In the data mining context an algorithm is based on one or more data mining techniques. That is some algorithms are based on decision trees, cluster analysis, or neural networks, or any other data mining technique. The goal(s) and tasks of the data mining process could be achieved using one or more algorithms.

In the following sections different data mining techniques will be discussed in details. Moreover, in chapter five the data mining techniques that will be used in this research will be evaluated and the chosen techniques will be justified.

4-7 Discussion of common data mining techniques

Any technique that helps extract patterns and hidden information from the database is called a data mining technique (Riedel, et al., 2000). Many classifications have been found for data mining techniques based on many variant factors (Ramakrishnan and Grama, 1999; Fayyad et al., 1996). These include:

1. *The induced representation* (decision trees, rules, correlations, deviations, trends, and associations);
2. *The data they operate on* (time series, discrete, labeled, continuous, or nominal);
3. *The application domains* (finance, economic, biology, Web log mining).

Adriaans and Zantinge (1996) proposed the following classification:

-Query tools;

-Visualization;

-On-line analytical processing (OLAP);

-Association rules;

-Cluster analysis;

-Decision trees;

-Statistical techniques;

-Genetic algorithms (GA);

-Artificial neural networks (ANN);

-Classification and regression techniques (CART).

An overview of these techniques is developed in the following sections.

4-7-1 Query tools

Traditional query tools are used first to analyze the data sets. Query tools are used to extract data that matches search criteria or to represent this data in a way that the user finds

easier to handle or interpret. By applying simple structured query language (SQL) the user can obtain a wealth of information. However, before we can apply more advanced pattern analysis algorithms, we need to know some basic aspects about the data set under study. With SQL we can uncover only shallow knowledge that is easily accessible from the data set: yet although we cannot find hidden knowledge. Adriaans and Zantinge (1996: 48) said, "*for the most part 80% of the interesting information can be abstracted from a database using SQL. The remaining 20% of hidden information requires more advanced techniques.*" However, for most organizations this 20% of hidden knowledge has 80% of the importance in relation to decision making, and 80% information volume represent only 20% in terms of value to the decision making process. The best way to start the search for shallow knowledge is to extract some simple statistical information from the data using SQL queries. For example:

- How many students in the university are taking the accounting major?
- What is the average GPA for the male students with an American diploma background?
- How many grants go to junior students?
- What is the nationality distribution of the students?

Decision makers can take decisions based on the output of the SQL statements. However, for the complicated decision situations that include many variables other tools are able to support the decision makers better than SQL.

4-7-2 Visualization

Visualization techniques depend strongly on the human side of the analysis (Berson, 1996). Data visualization is emerging as a technology that may allow organizations to process amounts of data and present it in a usable format. It is an interactive data manipulation technique that can process huge amounts of data. Colors, size, orientation, shape, and

behavior are used to present multiple dimensions. Using a graphical user interface visualization techniques provide non-computer users with the ability to navigate through data using their interpretations of the data displayed. Visualization techniques are valuable tools since they put the information we have in an easy and understandable way for both computer and non-computer aware people (Keim, et al., 1996), they also could be used throughout the data exploration process and are particularly useful during the initial stages of the high-level groupings of data sets.

The best set of rules or tables of data may reveal more information when visualized with color, relief, or texture in *2*, *3* or even *4 dimensions*. When using 4 dimensions, 3 dimensions are mapped onto the screen and the fourth can be expressed through the use of color (Chen and Paul, 2001; Berson and Smith, 1997). Object oriented multi-dimensional tool kits (e.g. Inventor) enable the user to explore three dimensions interactively; also advanced graphical techniques are used in virtual reality enable people to wander through an artificial data space. Some visualization software systems (e.g. SemNet) try to display the high dimensional structures and help the user grasp complex relationships (Chen and Paul, 2001). Complex visualization techniques are used in medical applications and also in computer games. For example visualizing scanned medical data to display the surface of an organ or bone (Johnson, et al., 1999). Foster, et al. (1999) introduced the concept of *distance visualization*. In distance visualization the environment is geographically distributed whereby data sources, end users, and visualization devices (i.e. computers and analysis tools) are scattered in different locations e.g. plants' images taken from the space.

For most users these advanced features are not accessible because they are expensive and require high computational power. As a result of that users have to rely on simple graphical display techniques that are contained in the query or data mining tools they are using.

However, even these simple methods can provide us with a wealth of information. A simple visualization technique that can be of great value is the *scatter diagram*; in this technique, information on two attributes is displayed in a Cartesian space. Scatter diagrams can be used to identify interesting sub-sets of the data sets so that we can focus on the rest of the data mining process (Adriaans and Zantinge, 1996).

4-7-3 On Line Analytical Processing (OLAP) tools

Although relational database management systems (RDBMS) are powerful solutions for a wide range of commercial and scientific applications, they are not designed to address the multidimensional information requirements of the modern business analyst, for example forecasting, and classification (Berson, 1996).

The key driver for the development of OLAP is to enable the multi-dimensional analysis (Pyle, 1999). Although all the required information can be formulated using relational database and accessed via SQL, the two dimensional relational model of data and SQL have some serious limitations for investigating complex real world problems. Also slow response time and SQL functionality are a source of problems (Berson, 1996). OLAP is a continuous and iterative process; an analyst can drill down to see much more details and then he can obtain answers to complex questions.

For example what is the relationship between majors, certificates, nationalities, and GPA? This question can be represented in 4-dimensions, see figure (4-6).

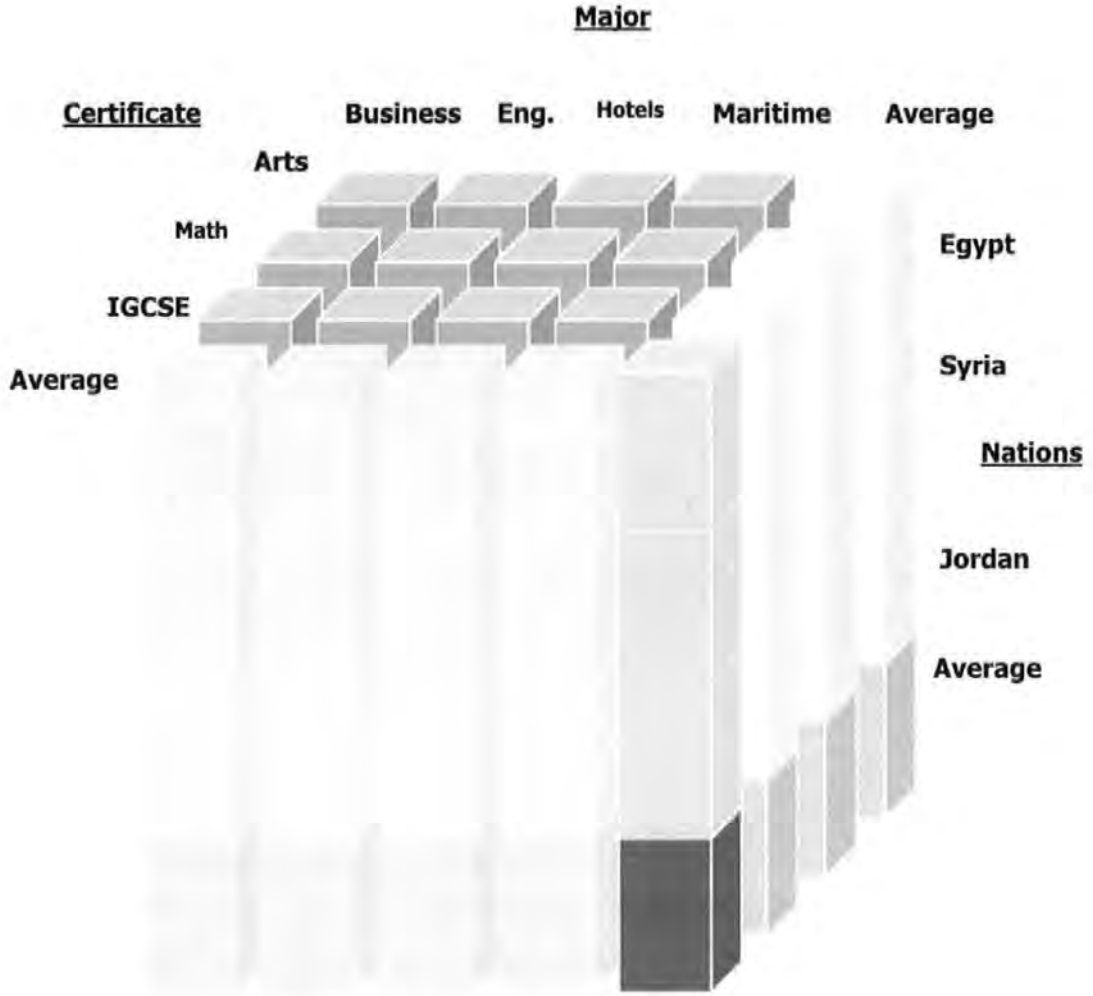


Figure (4-6). OLAP example.

Whilst multidimensionality is the core of a number of OLAP systems available (Pyle, 1999), there is a list of elements that determine which OLAP product to purchase:

1. *Multidimensional conceptual view*. The tool must support users with the level of dimensionality needed to enable the required analysis to be carried out;
2. *Transparency*. The heterogeneity of input data sources should be transparent to the users to prevent their productivity decreasing;
3. *Accessibility*. The OLAP system should only access the data required for analysis;
4. *Consistent reporting performance*. As the number of dimensions increases and the database size grows, users will expect the same level of performance;
5. *Client/Server architecture*. The OLAP system has to be compatible with the client/server architectural principles;

6. *Generic dimensionality.* Every data dimension should be in both its structure and operational capabilities;
7. *Multi-user support.* The OLAP system must be able to support a multi-user environment;
8. *Flexible reporting.* The ability to arrange rows, columns, and cells in a way that facilitates visual analysis.

Decision makers should prioritize the previous list elements to reflect their business needs.

Several researchers have stated that OLAP is an independent technique and is as powerful as the data mining process and techniques (Laudon and Laudon, 2001; Adriaans and Zantinge,1996; Fayyad, et al., 1996). Pyle (1999) stated that in every data mining application the analyst should expect to find some relationships between the variables that describe the data set. These expected relationships need confirmation and any OLAP tool can work well in either confirming or denying these relationships. Because of this, OLAP is one of the data mining techniques applied in the early stages of the data mining process. However, unlike other data mining techniques, OLAP does not learn and hence can not search for new solutions (Adriaans and Zantinge,1996; Berson,1996).

OLAP involves several basic analytical operations including consolidation, drill-down, and statistical techniques (; O'Brien, 1996):

1. *Consolidation.* Consolidation involves the aggregation of data, e.g. the total number of students at the university, total courses, and average GPA;
2. *Drill-down.* This is the opposite of consolidation and involves more detailed inspection of the underlying data, e.g. the break down of the total number of students into different nationalities that belong to the different majors with different GPA;
3. *Slicing and dicing.* Slicing and dicing refers to the ability to look at the database from different viewpoints.

4-7-4 Association rules

According to Wijsen, 2001; Liu, et al., 2000; Aas, et al., 1999; Berson and Smith, 1997; Adriaans and Zantinge, 1996; Agrawal, et al., 1996, the interest in discovering association rules from large relational tables has been increased recently. Association rules are focused on finding relationships (i.e. associations) between a certain attribute (i.e. target attribute) that the user is interested in, and the remaining attributes in a relational table. Association rules are often used as tools in DSS. Visualization techniques could be used for rules discovery.

The strength of association rules is that they can efficiently discover a complete set of associations that meet the user's requirements. However, there is no single algorithm that will automatically give the users everything of interest in the database. On the other hand, the major drawback of association rules is the large number of association rules produced, especially when the attributes are highly correlated, which makes it very difficult for to be analyzed and/or understood by a human being. In addition to that, finding all the rules is extremely computationally expensive. Moreover, some of the rules are found *useless*. Examples of useless rules are: the association between students' addresses and their corresponding majors (i.e. Students who live in Miami tend to have the Hotels major), the association between students' registration numbers and their corresponding GPA's (i.e. Students whose registration numbers are even tend to retain GPA > 2.0). To overcome these problems (i.e. large number of rules and useless rules) different scholars have proposed different measures including: *support*, *confidence*, *improvement*, *accuracy*, χ^2, and *coverage*. However, there is no consensus on neither which of those measures should be used nor the format of the rules. *The different viewpoints on the measures and the format of rules are explained and analyzed as follows*:

1. Wijsen (2001) said that the majority of work in association mining was focused on finding rules in the form of: $\forall t\left[(R(t) \wedge C(t)) \Rightarrow C'(t)\right]$, where t is a tuple (i.e. row or

record), C and C' are constraints that relate certain attribute values of tuple t with specified constants. The *pruning* process has to be undertaken to eliminate useless rules and keep those rules that are interesting to the users. Pruning is done using the level of support and the level of confidence. The *support* s of an association rule is the percentage of tuples satisfying both the left-hand and the right-hand side of the rule. The *confidence* is c if c% of the records satisfying the left-hand side of the rule also satisfies the right-hand side. Wijsen (2001: 437) said, "our notions of support and confidence not only have the same name, but also the same intention", in other words he did not differentiate between support and confidence. *Following is an example*:

[*Student* (t) $\wedge$ 80 $\leq$ t (high school percent) $\leq$ 100 $\wedge$ t (Nationality) = Egyptian] $\Rightarrow$ t (major) = Electronics, at 75% level of confidence. *This rule can be rewritten as follows*:

(High school percent: 80..100) $\wedge$ (Nationality: Egyptian) $\Rightarrow$ (major: Electronics), 75%

2. According to Liu, et al. (2000: 125), "Association rules are a fundamental class of patterns that exist in data". The objective of association rules is to find all associations between data that satisfies the user specified minimum support and minimum confidence. According to them, as association rule takes the form of $X \rightarrow Y \ni X \subset I$, $Y \subset I$, and $X \cap Y = \phi$; where Y is an item (i.e. value) of the target attribute, X is a set of items from the remaining attributes, I is a set of items (i.e. attribute value pair), and D is a set of data cases (i.e. records). The rule has support s if s% of the data cases in D contains $X \cup Y$. Liu, et al. did not use the confidence level in their study, they said that confidence level does not reflect the underlying relationships represented by the data. Based on this finding, they recommended the use of support, and chi-square test (χ^2) for independence and correlation as the basis for finding significant rules. Basically, χ^2 is a test that is used to determine whether or not to reject the notion that two variables are independent, the test is based on the comparison between the

observed frequencies and the corresponding expected frequencies; the closer the observed to the expected the greater is the weight of evidence in favor of independence.

3. Aas, et al. (1999) defined association rules as the co-occurrences. According to Aas, et al. rules are formalized in the form of: IF $\{A\}$ THEN $\{B\}$; where A and B are item sets and $A \cap B = \phi$. The IF-part is called the *antecedent* and the THEN-part is called the *consequent*. They associated each rule with a level of support and confidence. An item set has support s if s% of the transactions in the DB contains that item set (i.e. the probability that if an item is chosen at random will belong to that item set). Confidence is defined as: $Support(A \cap B) / Support(A)$. In order for the rule to be used for prediction, Aas, et al. recommended the use of as additional measure called *improvement*. Improvement is defined as: $Support(A \cap B) / Support(A)Support(B)$; if >1 the rule is useful and can be used for prediction. They emphasized that rules fail to be of interest to users if they match prior knowledge, and/or refer to uninteresting attributes.

4. Berson and Smith (1997) said that rules come in the form of "IF X, THEN Y". They mentioned that the left-hand side (i.e. X) is called the *antecedent*, and the right-hand side (Y) is called the *consequent*. The antecedent is consists of one or more conditions, whilst the consequent is just a single condition. Berson and Smith also said that rule induction does not imply *causality*. That is the left-hand side of the rule does not cause the right-hand side of the rule to appear. They also defined the interestingness of a rule by two measures; *accuracy* and *coverage*. Where accuracy refers to how often the rule is correct, whilst coverage refers to how often the rule applies. In other words, Berson and Smith replace the term confidence with accuracy, and support with coverage.

5. Adriaans and Zantinge (1996) emphasized that association rules help the managers describe and understand the data. Adriaans and Zantinge said that the mechanism through which valuable rules are identified is called *interestingness*; an interesting rule meets the minimum user requirements in terms of both *support* and *confidence*. They defined the support as the percentage of records that holds true each of the sides, whilst the confidence as the percentage of records that holds true for the right-hand side with all conditions in the left-hand side.

As a summary and For the purpose of this thesis, association rules are characterized by the following:

1. Associations between different attributes in a relational table;
2. The number of rules that can be found in any DB is almost infinite;
3. There is no single algorithm that is able to find all rules in a DB;
4. Rules do not mean causality;
5. Different forms for representing the rules were suggested, all of which can be rewritten in the IF $\{A\}$ THEN $\{B\}$ form;
6. Various measures for interestingness were proposed, however, the measures suggested by Aas, et al. are found complete and will be adopted because they added another measure for prediction (i.e. improvement). Wijsen is the only scholar to make no distinction between support and confidence, while in fact there is a difference. Also, Liu, et al. are the only scholars to suggest the use of χ^2 as a measure of interestingness. Besides the use of χ^2 means that the antecedent always represent one attribute.

4-7-5 Decision trees

A decision tree is a predictive model that provides a means of visualizing complex decision problems where the questions can be posed in sequence. The first question has a series of

answers and depending upon the answer given further questions and answers may follow. Each branch of the tree is a classification question, and the leaves are partitions of the data set with their classification (Berson and Smith, 1997). Another definition for the decision tree uses logical methods of describing regions of state. These logical methods could be interpreted in a "IF..THEN" rules space (Pyle, 1999). One variable is studied at a time in a decision tree. The start point is found when the variable that best classifies the data set is determined and the next stage follows. Another classifying variable and another splitting rule is found...etc. This process continues until some end criteria is found.

There are many algorithms for developing decision trees (Frasconi and Soda, 1999; Blockeel and De Raedt, 1997; Quinlan, 1986; Kononenko, et al., 1984; Patterson and Niblett, 1983). Quinlan's ID3 is the most acknowledged algorithm in the area of decision trees. The ID3 is an iterative algorithm that starts with a subset of the data called a *window*. The window is chosen at random and then the algorithm develops the tree which correctly classifies the subset into branches. Then the remaining data are classified using the tree. If each data record finds its correct classification then the process terminates. If not, a selection of the incorrectly classified data records are added to the window and the process continues (Quinlan, 1986).

The decision tree is a useful technique in both data mining and predictive modelling processes. It prevents the problems of *over fitting* (i.e. the algorithm searches in the limited data sets, so the algorithm might over fit the data) and does not require the user to specify how to handle missing data unlike other data mining techniques. The decision tree classifies the data into branches without loosing any of the data records.

Another fact about the decision trees states that the tree can be built with high degree of accuracy when using relational database management systems (Berson and Smith, 1997).

4-7-6 Artificial Neural Networks (ANN)

An ANN is a computer programs that implements complex pattern detection and machine learning algorithms to build predictive models from large database(s). In order for the ANN to detect patterns in the data sets, it should learn to detect these patterns and make predictions, in the same way a human does. ANN are widely used in many business applications. ANN are also of different types and are often used for cluster creation.

Increasingly vendors of ANN systems have added visualization formats to make the user able to understand and interpret the system's output.

Building ANN can be a time consuming process, however, when developed correctly they provide a very powerful predictive technique. Successful ANN require some data preprocessing, a good understanding of the problem, the target of the prediction and also the setting of some parameters that will drive its mission.

The advantage of ANN is the high predictive accuracy that can be applied to different problems. The disadvantages of ANN are the lack of ease of use and the difficulties of deployment (Berson and Smith, 1997).

4-7-7 Cluster analysis

Clustering is basically classifying unclassified data (Gordon, 1981; Everitt, 1980). The *data* to be classified consists of a set of *items* (sometimes referred to as *objects*, *fields*, or *records*). Each item is described by a set of characteristics called *variables* (sometimes referred to as *attributes*). The target of clustering is to classify the items in the data set into a number of *groups* (sometimes referred to as *classes*, or *clusters*), such that objects within one group have *similarities* with one another. Where the number of items is n, the maximum number of groups should be n-1 (i.e. the number of expected groups varies from 1 to n-1).

The clustering process is done in two steps; *firstly* to find the similarity or distance between each pair of items, *secondly* based on the similarity/distance matrix a clustering technique will be used to find out the number of groups and which items they contain. Finally the groups found are depicted in a graph called a *dendrogram*. There are many methods to find the similarity or distance between items, these methods are called *proximity measures*. All proximity measures end up with either a *similarity matrix* (if it is a similarity measure) or a *distance matrix* (if it is a distance measure). The similarity matrix is denoted by S_i, whilst the distance matrix is denoted by D_i. Whether it is S_i or D_i following rules hold true. $S_{ij} = S_{ji}$ (or alternatively $D_{ij} = D_{ji}$), and $S_{ii} = 1$ ($D_{ii} = 0$).

Based on these rules a distance or similarity matrix is *symmetrical* and thus only *half* of the matrix needs to be stored. In fact the leading diagonal is not stored and thus the proximity matrix occupies $n(n\text{-}1)/2$ storage locations.

Chapter five (Refer to section 5-7) discusses cluster analysis in more detail including; variable types, difficulties in measuring variables, proximity measures, and the various clustering techniques.

4-7-8 Genetic Algorithms (GA)

The term is a combination of both biology and computer disciplines, and sometimes referred to as simulated evolution. Berson and Smith (1997) said that GA refer to these simulated evolutionary systems, but more precisely these are the algorithms that dictate how populations of organisms should be formed, evaluated and modified.

In many ways genetic algorithms are close to the biological evolution, the analogy is summarized in the following table (4-1):

Biology	**Genetic Algorithms**
-Organism	Which is the computer program being optimized
-Population	The collection of programs undergoing simulated evolution
-Chromosome	The chromosome encodes the computer program
-Fitness	The calculation with which a program value is determined for survival of the fittest
-Gene	The basic building block of the chromosome that defines one particular feature of the simulated organism
-Locus	The location of the chromosome that contains a specific gene
-Allele	The value of the gene
-Mutation	The random change of the value of the gene
-Mating	The process by which two simulated programs swap pieces of them in a simulated crossover
-Selection	The best program is retained and the less successful are excluded by deleting them from computer memory.

Table (4-1). Genetic Algorithms and Biology.

Over time, these algorithms improve in their performance and as a result increase the efficiency of resolving problems. Genetic algorithms have been used to find optimal clusters based on a defined profit measure. By themselves GA can not detect outliers and do not create rules. They have been used to optimize the nearest neighbour classification systems for predicting sequences in time series (Berson and Smith, 1997).

4-7-9 Probabilistic graphical dependency technique

These models specify the probabilistic dependencies, which underlie a particular model using a graphical structure. The model specifies which variables are dependent on each other. These models are used for categorical or discrete-valued variables, however, some extensions also allow for the use of real-valued variables. Within artificial intelligence and statistics these models are built initially in the context of the probabilistic expert systems.

Although, graphical model induction is still not a mature discipline, it is of interest to the KDD applications since graphical forms of the model are easily understood by users (Fayyad et al., 1996).

4-8 Human-interactive KDD process

As we have seen, KDD is an integrated environment that enables the user to take decisions by carrying out the complex knowledge discovery process (Brachman and Anand, 1996). The output of the KDD process should be integrated with other front-end tools in the business environment, for example EIS and DSS. The user of these tools watches important events in the corporate data, or may be interested in discovering unknown patterns and hidden facts in the corporate or departmental database. The various roles of the user in the development process for each KDD application are:

- The user should understand the domain of the data as well as the specific analysis techniques to be able to utilize the discovered knowledge;
- The user of the KDD application should be able to understand the KDD techniques that have been applied and the data elements within the database;
- The user should also be able to understand the knowledge in specific problem architecture; otherwise the knowledge would be irrelevant;
- A certain class of users must be able to apply the discovered knowledge in the context of a real business application.

4-9 Example of the KDD process

For the rest of this chapter data from the Admission & Registration office at AASTMT will be used to clarify the concepts associated with KDD. Here we are interested in some particular data of importance to the decision makers who are involved with the admission & registration function. Answers to the following questions would be of value to the decision makers:

- How many students applied? And of them how many have been accepted?
- What is the nationality distribution of applicants?
- What is the minimum total mark accepted for entry at the AASTMT?
- What is the minimum percentage mark accepted at each department?
- How to visualize the nationalities, age, and departments in one 3-d format?
- How to predict the student profiles?
- To what extent do these student profiles reflect the AASTMT competitive position?
- What do the competitors' statistics say?
- Based on the statistics of students' nationality, what are the countries that need more marketing budgets?
- Etc.

To investigate these questions and to establish the KDD process, a sample of 1600 records were extracted from AASTMT application records and statistics (AASTMT application records, 1995; AASTMT statistics, 1995). Of these 1600, 1100 were accepted at AASTMT.

4-9-1 Data selection

The sample consists of the pre-stated 1600 records. Each record consists of serial number, application number, first name, nationality, sex, address, desire(s), percentage grade, accepted/ rejected, and department. In order to facilitate the KDD process a copy of this operational data is drawn and stored in a scratch database file[3], a sample records of this database is given in table (4-2).

SN	App.No	F.Name	Sex	Nation	City	Desire	% mark	A/R	Dept
1	697	Mary	f	Egy	Alex	Hot	95	A	Hot
4	1079	Asser	m	Egy	Alex	Bus	70	A	Hot
197	1484	Ismael	m	Syr	Dam	Eng	67	A	Eng
41	1570	Mohamed	m	Egy	Alex	Hot	59	R	
56	19	Lamees	f	Pal	Dub	Bus	53	R	

Table (4-2). Sample of the original data.

[3] MS-Access was used by the researcher for this purpose.

-*Note*: SN stands for serial number, App.No for application number, F.Name for first name, Nation for nationality, A for accepted & R for rejected, and dept for department. Also, the departments, and nationality names have been curtailed for space reasons in this table.

4-9-2 Cleaning

Several methods are available to clean the data i.e. remove errors. Some of these methods can be executed in advance while others are only invoked after errors are detected at the coding or the discovery stage (Adriaans and Zantinge,1996). There are three different sources of errors; *duplication*, *domain inconsistency*, and *missing values*.

A very important element in a cleaning operation is the *de-duplication* of records (table 4-3). In the student database file a student may be represented by more than one record.

SN	App.No	Name	Sex	Nation	City	Desire	% mark	A/R	Dept
270	316	Ehab	M	Egy	Alex	Eng	70	A	Eng
271	623	Ehab	M	Egy	Alex	Eng	70	A	Eng

Table (4-3). De-duplication of records.

In the present example we have two different records for the same student. This may be due to two people submitting the same student data without the student being aware of that. Of course, we can never be sure of this, but de-duplication algorithm using analysis techniques could identity the situation and present it to a user to make a decision.

There are also cases in which people spell their names incorrectly or give incorrect information about themselves by slightly misspelling their name or by giving a false address. The data cleaning process affects the quality of the mining process, because if the process of data cleaning was performed thoroughly, the results of data mining would be helpful and trustworthy.

The second type of data errors that frequently occurs *is the lack of domain consistency*, (table 4-4).

SN	App.No	Name	Sex	Nation	City	Desire	% mark	A/R	Dept
541	350	Ahmed	M	Egy	Cairo	Eng	70	A	
215	476	Ahmed	M	Alex	Lyb	Eng	70	A	Eng

Table (4-4). Domain consistency.

Notice that in table (4-4) the first student department is empty, however, this attribute should have a value. Empty values are a source of problems to the data mining process, because this might affect the type of patters discovered. If the data item is not defined it should be NULL. In the second record there is an inconsistency in the domain of nationalities and cities, and replacement would happen between city, and nation. If data is unknown it should be represented as such in the database. In our example, we have replaced part of the data with NULL values and corrected other domain inconsistencies. The result is shown in table (4-5).

SN	App.No	Name	Sex	Nation	City	Desire	% mark	A/R	Dept
541	350	Ahmed	M	Egy	Cairo	Eng	70	A	NULL
215	476	Ahmed	M	Lyb	Alex	Eng	70	A	Eng

Table (4-5). Domain consistency-1.

4-9-3 Enrichment

In our present example, assume that we have some extra data about the students' family annual income, and the student secondary school. Decision makers who to set the tuition fees rates or determine certain marketing policies could use this extra data. Enrichment is carried out wherever it is possible to get or buy extra relevant data. Table (4-6) describes enriched students' records.

SN	App.No	Name	Income	School	Nation	Dept
100	801	Ola	200.000	Saudi Sch.	Sau	NULL
200	802	Alaa	350000	IGCSE	Leb	Eng

Table (4-6). Enrichment.

4-9-4 Coding (Pre-coded data)

In the next stage, we select only those records that have enough data to be of value. In table (4-7) extra data has been added to the original data.

SN	App.No	Name	Inc.	Car	Sex	Nation	City	Desire	%	A/R	Dept
911	350	Tarek	Null	Null	m	Egy	cairo	Eng	60	A	Null
111	714	Wael	150	Null	m	Null	Alex	Mar	85	A	Eng

Table (4-7). Enriched table.

A general rule states that any deletion of data must be a conscious decision, and only taken after a thorough analysis of the possible consequences. However, in some cases lack of data can be a valuable indication of interesting patterns.

In the cases presented in table (4-7) for the students Tarek, and Wael, we lack some data concerning them, so we choose to exclude their records from the final sample. But excluding the records does not mean deletion, because the deletion decision is always questionable due to any causal connections.

Adriaans and Zantnige (1996) stated that it is better to delete incomplete data instead of getting incorrect results. However, due to the deletion consequences it is better to exclude rather than to delete. Deletion disables any possibility for further analysis to be carried out when the incomplete data fields are being filled in.

Next we carry out a *selection* of the fields (i.e. columns or attributes). In our example we are not interested in the students' name, so their names are removed from the sample. Up to this point, the coding phase consisted of nothing more than simple SQL operations but now we are entering the stage where we will perform some data transformations. By this time, the data records in our database are much too detailed to be used as input for pattern recognition algorithms (i.e. there is a negative relationship between the database size and the algorithm performance). For example, department, nationalities, and cities represent a complex set of data that has to be coded before used as inputs otherwise the performance of the algorithms will be very slow.

4-9-4-1 Coding (post-coding data)

Coding is an operation that occurs frequently in the KDD context and is carried out to improve the performance of the algorithms used in analyzing the data set.

For example, income could be transferred into categories as follows:

- Annual income less than 100,000 L.E is called *starting*, coded 01;
- Annual income from 100,000 to less than 200,000 L.E is called *moderate*, coded 02;
- Annual income from 2,00,000 to less than 5,00,000 L.E is *premium* income, coded 03;.
- Annual income more than 5,00,000 L.E is called *super premium* income, coded 04.

Nationality is also coded, so instead of having one attribute with 20 different possible values, we create twenty codes; *the code consists of two digits,* one for each nationality. "10" means that the student is Egyptian, 20 for Sudanese, 30 for Libyan etc. Such an operation is called *Flattening*; an attribute with *cardinality n* is replaced by *n* codes. Applying the concept Flattening to the Sex results in 01, 02, which means that male is replaced by 01, and female is replaced by 02 *(cardinality 2)*. Table (4-8) represents the records that resulted from the coding process.

SN	App.No	Inc.	Car	Sex	Nation	City	Desire	%	A/R	Dept
911	350	02	Null	01	10	101	1	60	A	1
111	714	01	Null	01	30	301	4	85	A	4

Table (4-8). The coding effect.

Using this coding we can obtain the following more rapidly:

1. The number of Egyptian students in the Maritime department who have a premium income. OR;
2. Which Syrian students with Engineering major have an average mark above 60%.

Referring to section (4-4) in this chapter, which at the end states that the DW will enhance the KDD process. A typical DW will be several *Gigabytes* or *Terabytes* in size and contain millions of records so the coding process will enhance its performance. Another coding advantage is the resources savings (i.e. physical disk space, cost, and performance).

4-9-5 Data Mining Techniques

In section (4-7) of this chapter the data mining techniques were discussed with respect to many classification mechanisms. However, the KDD example here will not use all of the techniques for many reasons:

- There is no single application that uses all of the data mining techniques, because of the differences in the goals of the application domains;
- Time and effort constraint;
- The use of the techniques, which are relevant to this dissertation, will be justified in chapter five (Refer to section 5-4).

4-9-5-1 Traditional query tools

A good way to start is to extract some simple statistical information from the data set. *Table (4-9) provides statistics on students.*

Department	Private*	Sponsor**	Transfer	Total
Nautical	54	6	0	**60**
Maritime Eng.	37	2	0	**39**
Mechanical Eng.	51	0	2	**53**
Computer Eng.	93	6	6	**105**
Power Eng.	64	0	0	**64**
Electronics Eng.	78	5	5	**88**
Construction Eng.	61	0	0	**61**
Managerial Eng.	58	0	0	**58**
Business Adm.	181	6	7	**194**
Hotels & tourism	27	0	1	**28**
Grand total	**704**	**25**	**21**	**750**

***Private** Private student means that he/she pays for himself/herself.
****Sponsor** Sponsored student appears when a certain agency/authority pays for the student fees and accommodation, regardless of the reason.

Table (4-9). Statistics.

-*Note*: 350 students should be added to the grand total of the previous table to give the total of 1100 accepted students. They represent the number of students that will join the preparatory program so that, after succeeding, they can join next semester.

From table (4-9), and the original data sets, shallow knowledge can be extracted and represented. An example of this is shown in figure (4-7), from which we can see that the percentage of sponsored students at the AASTMT is about 3%, with another 3% being transfer students. Thus "*if the AASTMT has 1000 students then 940 are predicted to be private, 30 are transfer, and 30 are sponsored*".

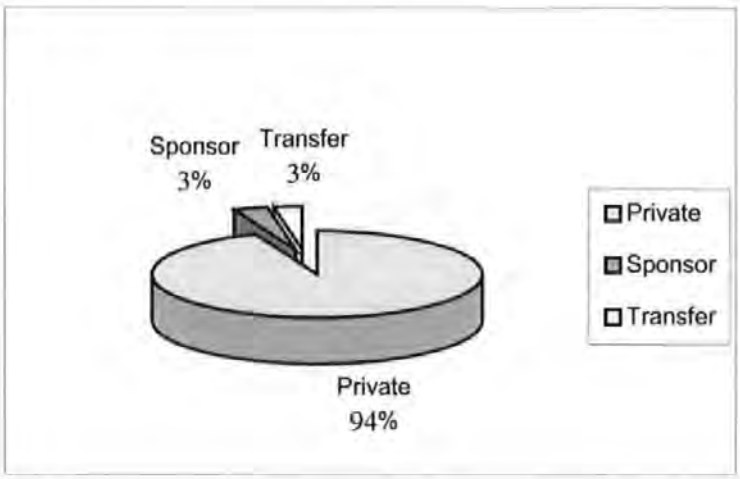


Figure (4-7). Student sponsorship rate.

The analysis is not a goal in itself, rather the implications of the analysis to the executive who will take a decision according to that analysis. For example, the drill down capability reveals that the 3% of students who are sponsored historically joined the Nautical department, which currently has a little demand (AASTMT statistics, 1990-1997). May be the reason is the student is actually sponsored but claims the opposite to save money, since sponsored students pay higher fees. Figure (4-8), shows the student distribution rates among departments.

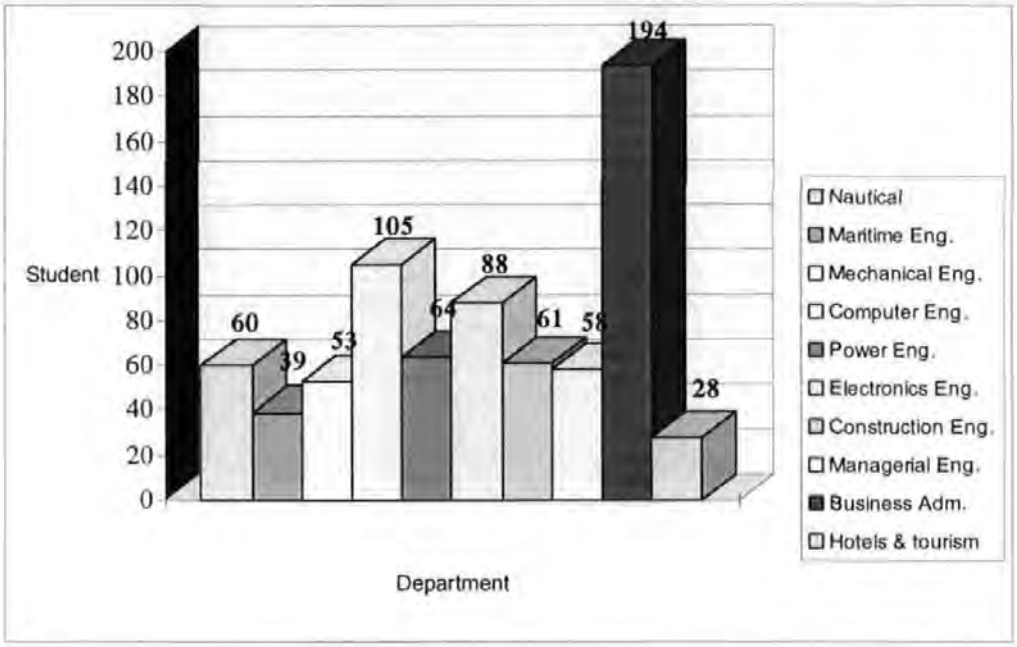


Figure (4-8). Department distribution.

Figure (4-8) also indicates that the highest number of students joined the Business Administration department. An executive should also examine all the implications associated with this fact.

Data mining presents the results of the analysis in a format that is readily understandable to the executives. However, a further feature of data mining is that it allows the executive to analyze issues in greater depth using the drill down capability.

For example, the fact that Business Administration has the highest demand could be investigated further using the mining techniques to provide possible explanations such as:

- There is a trend in the labor market to give the Business graduates greater number of work opportunities;
- Competitor Universities are weak in this field;
- The excellent facilities the AASTMT has made this possible;
- The staff members of the Business Administration department are superior to its rivals, and use relevant and up to date textbooks and literature sources;

- The marketing efforts of the AASTMT are biased to the Business Administration department;
- The reason of this is that the Business Administration graduates excel at their work.

Regardless of what the reason is, the drill-down capability is an important tool for executives, and this is the core benefit that data mining offers Executive managers.

Also the analytical ability can reveal hidden knowledge. For example the executive may want to track the transfer student information. Figure (4-9) provides a good example.

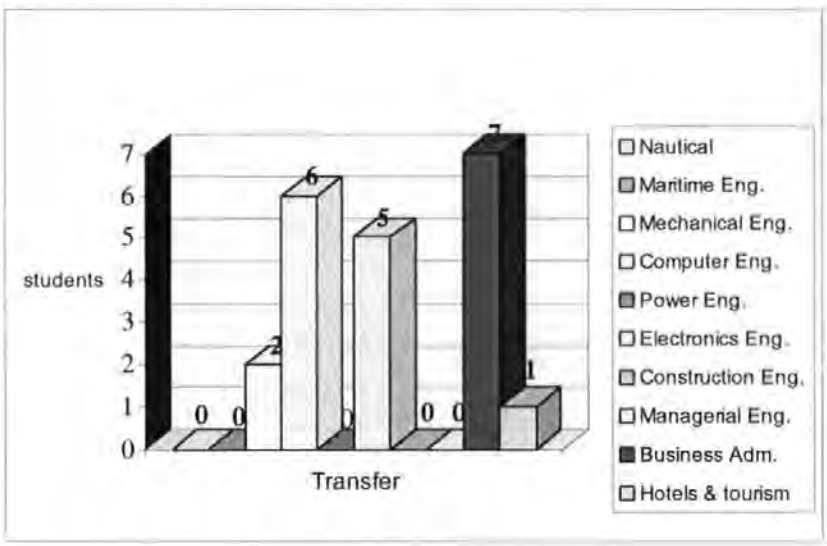


Figure (4-9). Transfer students.

Data mining can make it easy for the executive to analyze the results. The graph shows that some of the Engineering departments (Marine, Power, Construction, Managerial) and the Nautical have a 0% transfer. This might be due to the following:

- Students of the other departments see no value of applying to these departments;
- Course structure makes it difficult for students to transfer to them;
- A decrease on student GPA if transferred to them, due to the cancellation of some courses he has completed that will not be counted;
- These departments do not accept transfer students.

The analytical ability offers the different possibilities to the executives and enables hidden knowledge and patterns to be found. The role of the executive is to take corrective actions.

4-9-5-2 Visualization techniques

In figure (4-10), a visualization technique plays an important role as it puts the information in a format that is easier to understand. Table (4-9) has been transformed to the following figure (4-10).

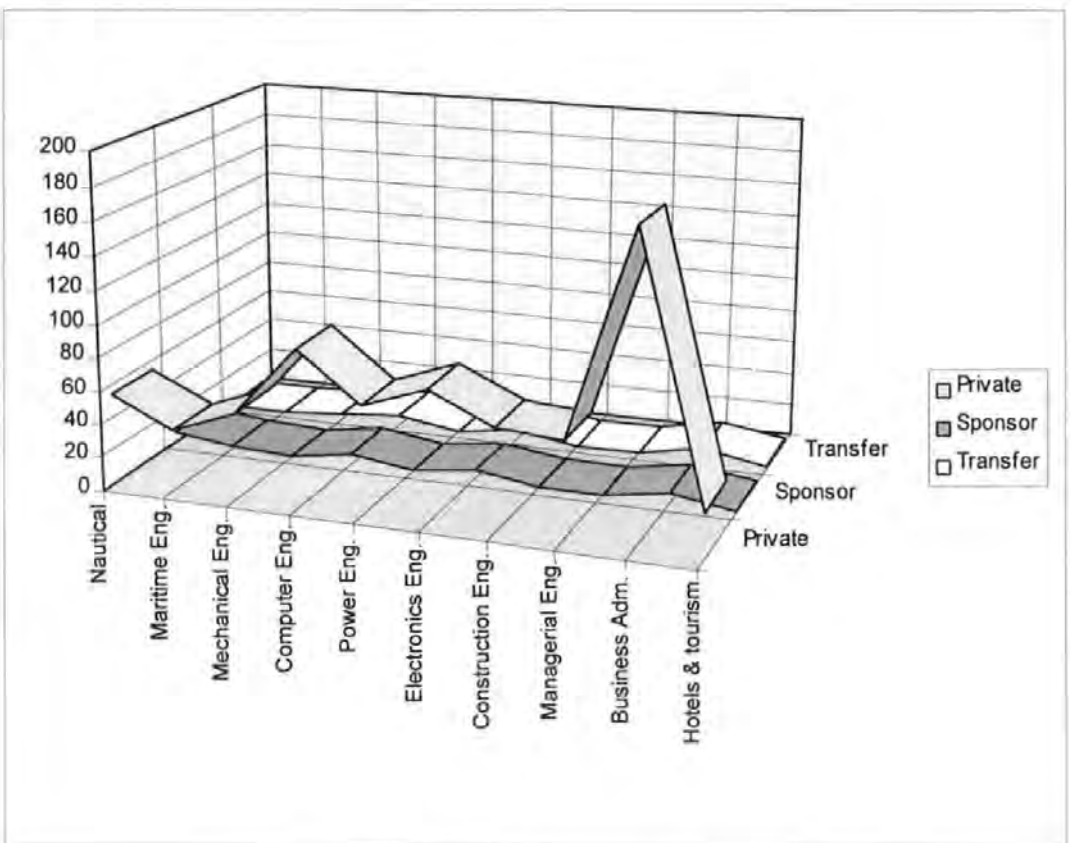


Figure (4-10). Visualizing new Knowledge.

Figure (4-10) is a three-dimensional chart that will enable the executives to reach a multi-dimensional knowledge in an easy way. However, different executive managers might have different interpretations. For example:

- The number of private students is higher than sponsored and transfer students;
- The Business Administration has the largest number of private students.

4-9-5-3 OLAP

We can plot table (4-9) in the following format in figure (4-11). The figure represents the number of students within the different departments, regarding if they are sponsored, private, or transfer students, but in detailed categorical levels.

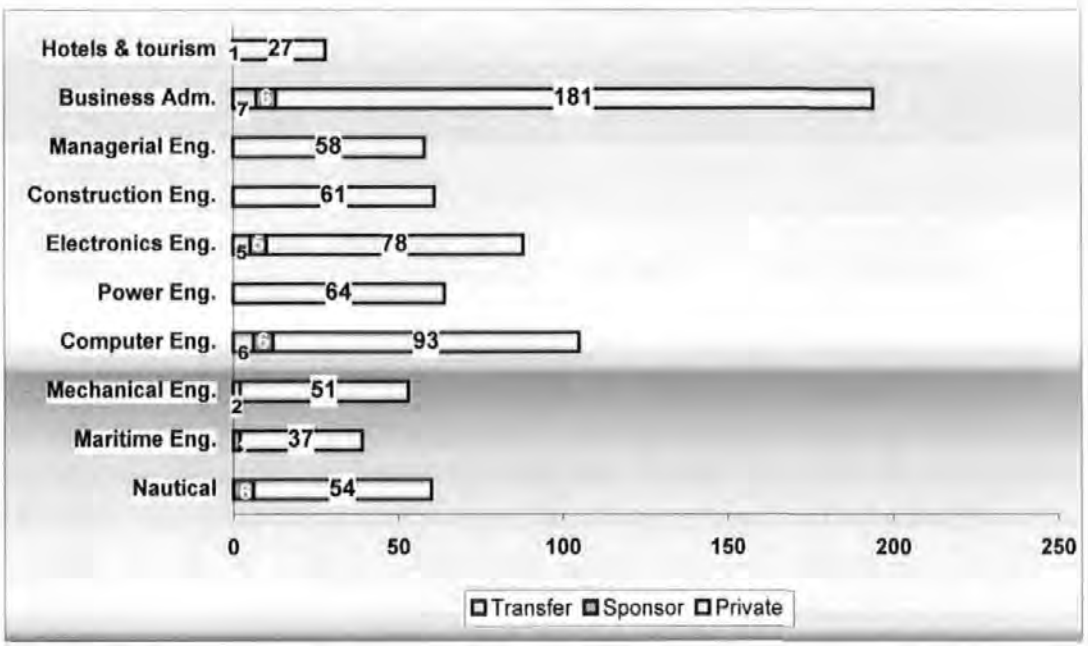


Figure (4-11). Slicing and dicing.

4-9-5-4 Association rules

For the purpose of the data set under study many rules could be found. *For example*:

IF {Department= Business Administration AND Private= Y} THEN {Nationality= EGY}

$S(A) = 181/750 = 24\%$ $S(B) = 600/750 = 80\%$

The two sides have co-occurred 150 times concurrently; $S(A \cap B) = 150/750 = 20\%$

$C = 20\%/24\% = 83\%$

Improvement = 20%/(24%*80%) = 1.04, which indicate that the rule is useful and can be used for prediction.

Various decision makers could interpret the previous rule differently. Based upon their interpretations, they can derive knowledge that would enable them to better understand the students' records and to make decisions. *The previous rule could be interpreted as follows*:

- The majority of the students in the Business Administration department are Egyptians;
- The majority of the Egyptian students in the Business Administration are not sponsored;
- Setting the Admission requirements for the Business Administration department should take into consideration the Egyptian Students' profiles as they represent the majority;
- Extend the marketing efforts and budgets to recruit students from other nationalities;
- Setting the fees for the Business Administration department should take into account the average income for the Egyptian students' families;
- Understanding the students' profiles in the Business Administration department;
- Understanding the Egyptian nationality distribution amongst the departments.

4-9-5-5 Cluster analysis[4]

Cluster analysis groups the data in two steps; firstly to find out similarities/distance between the data items, secondly to use a suitable clustering technique to find out each item's group. In the following example five students will be classified into groups based on their values in two variables.

Students	Mark	Major
Student 1	62	1
Student 2	65	1
Student 3	92	4
Student 4	85	4
Student 5	70	9

Table (4-10). Sample data set.

Following is the sample characteristics that is required for some similarity/distance measures:

[4] For details on cluster analysis in terms of similarity measures and clustering techniques refer to chapter five section (5-7).

Characteristics	Var 1	Var 2
Range	30	8
StdDev	13.07	3.27

Table (4-11). Sample characteristics.

Following is the distance between the five items based on a modified Euclidean metric measure.

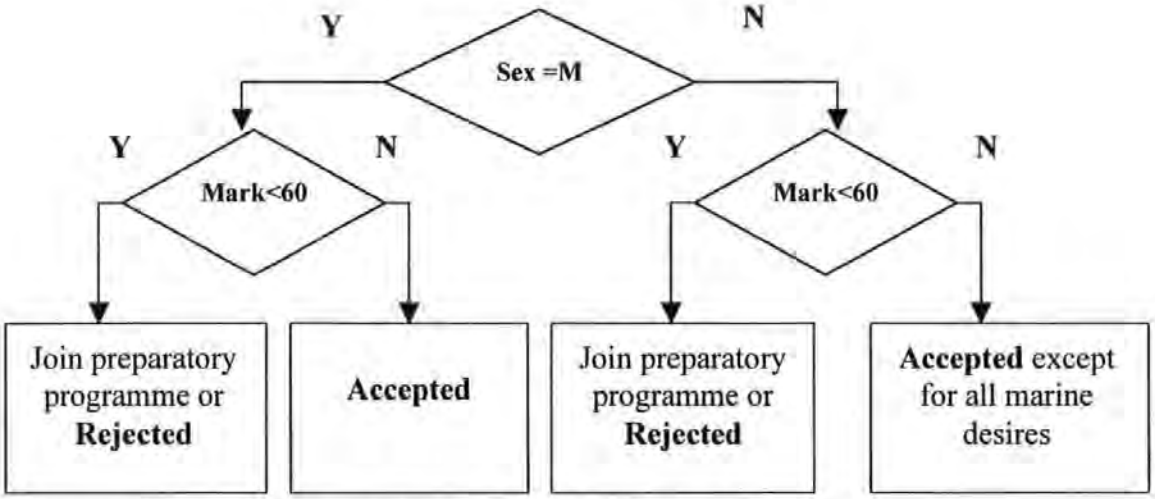
$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ .23 & 0 & & & \\ 2.47 & 2.26 & 0 & & \\ 1.98 & 1.78 & .53 & 0 & \\ 2.52 & 2.47 & 2.27 & 1.91 & 0 \end{pmatrix} \end{array}$$

Using the nearest neighbour (NN) hierarchical clustering technique, based on the modified Euclidean metric measure $\mathbf{D_1}$: item (i.e. student) **1** joins item **2** to formulate the *first group*, in the second fusion item **3** join **4** to formulate the *second group*, in the third fusion the *first group* joins the *second* to formulate *group three*, finally item **5** joins the *third group*.Clustering is useful for classification and prediction pruposes.

4-9-5-6 Decision trees[5]

The following decision tree is found in the admission and registration data set under study. The tree is based on the following attributes; Sex, Mark, and the decision to accept or reject an applicant. The tree was built using SQL statements.

[5] More informative decision trees could have been built if more attributes were available in the data set under study e.g. age, type of high school certificate, graudation grade.



- The tree follows the accept/reject decision logic at the AASTMT.

Rule 1: IF Sex <> M AND Mark < 60, THEN an applicant joins preparatory programme OR Rejected.

Rule 2: IF Sex <> M AND Mark > 60, THEN an applicant is Accepted except for all marine desires.

Rule 3: IF Sex = M AND Mark < 60, THEN an applicant joins preparatory programme OR Rejected.

Rule 4: IF Sex = M AND Mark > 60, THEN an applicant is Accepted.

Figure (4-12). Decision tree for the Accept/Reject decision.

4-9-6 Reporting

Reporting the results of data mining can take many forms. The illustration already described in this chapter has given a good idea of the options available. *In general, one can use any report writer or graphical tool to make the results of the process accessible* (Adriaans and Zantinge, 1996). The material in this chapter has also provided a good indication of the interactive character of the KDD process: there is a consistent interplay between the selection of data, cleaning, data mining, and the reporting results.

4-10 Research and application challenges for KDD

1. *Data mining and data security.* Since the data mining techniques provides aggregation and drill down of information based on data from different data sources, it might invoke someone's personal data (i.e. security violation). Little research efforts have been directed towards addressing the security of data mining techniques (Agrawal and Srikant, 2000);

2. *Larger databases/algorithm scalability.* The data mining techniques used as steps in the KDD process should be able to handle database with hundreds of tables and fields, as well as millions of records. These databases normally are of gigabytes and sometimes of terabytes in size. The techniques should be efficient in dealing with these volumes of data. Solutions might be the sampling or approximation methods, or massive parallel processing (Fayyad, et al., 1996);
3. *High dimensionality.* This problem is closely related to the very large number of fields/variables in the database, so the dimensionality is high. The high dimensionality creates a larger search space, which causes the model induction to be complex and time-consuming. This also will increase the probability that the data mining technique will find spurious patterns. The solution might be to reduce the effective dimensionality of the problem and the use of prior knowledge to remove the irrelevant variables;
4. *Overfitting.* This happens when the mining algorithm searches for the best estimates of the parameters for one model using a limited data set, which may lead to an overfit of the data and the models having poor performance. Possible solutions might be cross-validation, regularization strategies, and updating the parameter estimates;
5. *Changing data and knowledge.* Data collected from rapidly changing environment may cause the discovered patterns to be irrelevant and spurious. An example is stock market data. Possible solutions might to update the patterns and treating the change as an opportunity for discovery;
6. *Missing and noisy data.* Here the problem occurs when important data items are missing. Some inductive logic programming techniques (ILP) can resolve this problem. IPL is a machine learning language where relations are represented in a deductive database context. A deductive DB is capable of deducing information by applying inferential rules that are stored in the DB (Date, 1995). Basic ILP techniques include

relatively least generalizations, inverse resolution, searching refinement graphs, and use rule models (Dzeroski, 1996);

7. *Complex relationships between fields.* Data mining algorithms must be able to utilize hierarchical attributes or any other sort of related attributes. Traditionally mining techniques have been developed to deal with simple attribute-value records. Dzeroski (1996) said that new techniques are being developed to handle the relations between variables in the area of ILP;
8. *The business-Data mining gap.* There has been a gap between an organisation's business objectives and the parameters by which the mining techniques should be chosen. Little research has been addressed to find a technique or method by which there would be a link between the objectives and the chosen mining techniques (Ali and Wallace, 1997);
9. *Understandability of discovered knowledge.* One critical factor in KDD applications is *to* make the output of the systems understandable by the decision maker. Possible aids include graphical representation of the knowledge discovered (e.g. visualization techniques);
10. *User interaction and prior knowledge.* Many of the recent KDD tools are not interactive and can only incorporate prior knowledge about the problem in a simple way. Bayesian approaches use prior probabilities over data and distribution as one source of encoding prior knowledge (Fayyad, et al., 1996).

Chapter summary

- There are four different types of knowledge; shallow, multi-dimensional, hidden, and deep knowledge.
- KDD is the process of fining hidden knowledge, patterns and unknown facts from the data sets.
- KDD has been used successfully in many disciplines.
- Data mining is a step in the KDD process.

- KDD consists of these steps: developing an understanding of the application domain, creating a target data set, data cleaning and preprocessing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm(s), data mining, interpreting the information gained by the mining techniques, and finally consolidating the discovered knowledge.
- The goals of the data mining techniques are prediction and description.
- The data mining tasks are clustering, classification, summarization, dependency, regression, and change detection.
- Data mining techniques are many. The techniques used for different tasks up to the goal of the KDD process. Techniques are query tools, visualization, on-line analytical processing (OLAP), association rules, cluster analysis, decision trees, statistical techniques, genetic algorithms (GA), artificial neural networks (ANN) and classification and regression techniques (CART).
- During the KDD process, the user should be able to interact with the application and that necessitate that he should understand the goal of the application and the techniques used.
- KDD as an emerging field facing many challenges. For example larger databases/algorithm scalability, high dimensionality, overfitting, assessing statistical significance, changing data and knowledge, missing and noisy data, complex relationships between fields, understandability of discovered knowledge, and user interaction and prior knowledge.
- The KDD process especially the data mining techniques will be enhanced when used with the data warehouse (DW). The DW needs a front-end tool to utilize its information like DSS and EIS. The next chapter will introduce these components together and describe the methodology that will be employed in this context.

Chapter five

The blend of DSS, DW, and KDD

This chapter will establish the relationship between DSS, DW, and KDD, as they constitute the three corner stones of this research and examine how these components work together in the proposed DSS methodology. A new DSS definition is proposed to cope with the proposed methodology. The tools and mechanisms of the proposed DSS methodology are explored together with the KDD techniques that will be employed in this research. Justification of the tools and mechanisms employed to complete the practical implementation of the DSS is also introduced including, the Cool: Gen 5.0 CASE tools and MS-SQL Server.

5-1 The relationship between DSS, DW, and KDD

In order to produce effective decision support systems that are able to enhance the quality of the decision making process research efforts have concentrated on introducing different combinations of three components (i.e. DSS, DW, and KDD), whilst other research efforts have ascertained the importance of linking them together. Moreover, this research has found that there have been little effort made to integrate these three components in certain application areas or to explain how these components work together and what tools and mechanisms used.

The research efforts can be classified into two groups:

-The first group includes those researchers who have tried to show the importance of linking two components together:

> -Inmon and Hackathorn (1994: ix) said, "The primary application that is served by the data warehouse is the DSS. The data warehouse provides the foundation needed for effective DSS processing. It is true that the DSS can be built without the DW, but the effective DSS can not be built without the DW."
>
> -Mattison (1997: 25) said, "The history and much experience-based analysis of data warehousing have shown us unequivocally that the deployment of successful data warehouse and data mining initiatives, in any industry, occurs

only when the system that is being implemented is derived from and delivered to the business in a way the is industry specific and tailored specifically to the needs of the business for which it is being created."

-Devlin (1997: 37) said, "IS managers can thus point to significant maintenance and data management problems resulting from this approach (i.e. traditional development approaches) to providing decision support. The solution is, of course, implementation of a data warehouse."

-Barquin (1997: 11) said, "A data warehouse without a DSS/EIS front-end to assist end-users in making decisions about the business of the business, is probably an unsuccessful data warehouse."

-Turban and Aronson (1998: 123) said, "A data warehouse provides data that are already transformed and summarized, therefore making it an appropriate environment for more efficient DSS and EIS application development and access."

-Burn-Thornton, et al. (1998: 2) said, "It has been previously proposed that data mining techniques have the potential to provide the underpinning technology for decision support systems."

-Humphries, et al. (1999: 263) said, "Data mining tools will continue to mature, and more organisations will adopt this type of warehousing technology. Learning from data mining applications will become more widely available in the trade press and other commercial applications, thereby increasing the chances of data mining success of late adopters."

-The second group of researchers have assured the importance of linking the three in future research:

-Adriaans and Zantinge (1996: 1) said, "The combination of data warehousing, decision support, and data mining indicated an innovative approach and totally new approach to information management."

-Berson and Smith (1997: 120) said, "The principal purpose of data warehousing is to provide information for strategic decision making. The user interacts with the data warehouse using front-end tools. These tools include; data query and reporting tools, application development tools, EIS tools, OLAP tools, and data mining tools."

-Taha, et al. (1997: 77) said, "The fields of DSS, OLAP, data mining, and DW are interrelated to provide good solutions for meeting certain customer needs, in particular, the need to mine a huge amount of well-organized data and the analysis of data for use by decision makers. The organisation DSS has a DW as data repository and a data mine where the data is organized for the purpose of information retrieval and analysis."

-Turban and Aronson (1998: 135) said, "Organisations are recognizing that their data contain a gold mine of information if they can dig it out. Consequently, they are warehousing and data mining for users to obtain information on their own and to establish relationships that were previously unknown."

-Han, et al. (1999: 46) said, "Current data warehouse systems have provided a fertile ground for systematic development of this multidimensional data mining."

-Gray and Watson (1999: 1) said, "Data warehouses, OLAP, and KDD are leading to new ways of performing decision support systems and creating executive information systems for data rich environments. Yet, these developments have received almost no attention from academics either in their research on in teaching."

-Cooper, et al. (2000: 566) said, "DW and other advances in IT are now solving the very difficult technical problems. They make it possible to organize, store, and retrieve huge volumes of information and to select critical information for a given decision. However, before organizations can realize that "grand promise" of MIS, most will have to reshape their business processes and decision making cultures to take advantage of the technology's new capabilities. This is a non-trivial transformation."

The contribution of this research is in integrating the three components DSS, DW, and KDD together and developing an integrated system for use by decision makers in the area of admission and registration in universities.

5-2 The proposed DSS definition

Organisations are currently faced with strong and increasing competition. Thus for managers to be effective and make sound decisions they need an IS that is able to process a broad spectrum of data, both internal and external, and also has the capability to perform in-depth searches within this wealth of historical data. The IS should, where appropriate, enable managers to look at the organisation's historical data in order to reach the deep knowledge and thus be able to make more informed decisions.

Because of the quantity of data an organization owns, and the daily increase in the volume of the data the IS must be able to handle large dynamic data sets. The rapid growth of the World Wide Web (*WWW*) is a good example to be considered here. According to Beeri, et al. (1998) every day hundreds of gigabytes of data are distributed around the world, and it is no longer possible to monitor this increasingly rapid development. However, according to recent statistics (2001) from LINX (London Internet Exchange) their Server handles 5 Gbits per second on average IN, and 5 Gbits per second on average OUT.

Adriaans and Zantinge (1996: 2) said, "We are confronted with the new paradox of the growth of data, *that more data means less information"*. The mechanical production and reproduction of data forces us to adapt our strategies and develop mechanical methods for filtering, selecting and interpreting data. Organizations that excel in doing this will have a better chance of surviving and because of this, information itself has become a production factor of importance.

Unfortunately, the traditional DSS definitions presented in chapter two do not provide a comprehensive view through which DSS would be able to respond to the new organisation needs. Traditional DSS submit the shallow knowledge to the user manager. The user manager will require DSS that respond to the new needs and are able to handle the different types of knowledge especially deep knowledge. The KDD process has make it possible for user managers to extract all types of knowledge in an efficient and rapid manner that has not been introduced before (Taha, et al., 1997; Adriaans and Zantinge, 1996).

Due to the shortcomings of the traditional DSS definitions *(refer to chapter two, item 2-5-1)*, and responding to the organisations' knowledge requirements, a new DSS definition will be introduced as follows.

> "A *DSS is a computer-based information system that deals with semi-structured and unstructured problems facing managers at all management levels. The DSS goal is to enhance the decision quality and the manager effectiveness. To do so, the DSS integrates itself to the strategic data store which is the data warehouse (DW), and to the knowledge discovery in database (KDD) process that will find the deep knowledge and hidden patterns in the DW and present them to the DSS user.*"

The definition emphasizes many issues, these are:

- DSS are CBIS;

- DSS can be used by managers at all the management levels;
- DSS deal with semi or unstructured problem types;
- DSS aim at enhancing the manager's effectiveness and the decision quality;
- DSS are linked to the organisation databases through the data warehouse;
- DSS target the organisation's deep knowledge through utilizing the data mining techniques.

5-3 Data mining techniques

The role of data mining techniques is to enable the user managers to accomplish tasks that cannot be performed in a traditional way (e.g. *normal SQL search).* For example, the following type of questions can be answered easily without the need for data mining techniques: What is the average age of the students at AASTMT? What is the average GPA at the College of management and technology? How many Syrian students took the Hotels and Tourism as their specialization from 85 till 97? How many nationalities are available at AASTMT?

However, many of the tasks performed by the user manager can not be supported solely using the normal search statements. For example, the following questions can be easily targeted and answered using data mining techniques: What are the Egyptian students' profiles at AASTMT? How could we visualize age, against nationality and GPA? Is their any relationship between the freshmen students' certificates and their GPA after applying to AASTMT? How can we build course forecasts based on the historical data we have? What is the nationality distribution across years on different majors from various secondary certificates? Etc.

Data mining will add value to the historical data extracted from the DW by finding new trends and discovering new knowledge. *The following are the enhancements of data mining to DSS implementation*:

1. Finding the hidden knowledge, unexpected patterns, and new rules from large databases;
2. Responding quickly and easily to the complicated questions that would have take days or even months to be answered manually;
3. Enabling predictions to be made;
4. Data mining is a technique that helps extract more information from the available data;
5. Different data mining techniques provide the executive manager with the ability to reach the organisation's deep knowledge.

The data mining techniques that will be used in this research are chosen based on the following factors:

1. To match the business application we are studying which is the admission and registration function in universities;
2. To match the goals of the mining process for this business application. These goals are both description and prediction i.e. we need to describe the students' data and also to use the existing data to make future predictions;
3. The data mining tasks selected to meet the data mining goals of this business application are summarization, and clustering.
4. To match the technique characteristics which were found in literature of the previous studies. See the next section.

5-4 The Data mining techniques chosen for this research

The following table (5-1) is based on the work of Chen and Paul, 2001; Wijsen, 2001; Guha, et. al, 2000; Liu, et al., 2000; Aas, et al., 1999; Foster, et al., 1999; Ramakrishnan and Grama, 1999; Pyle, 1999; Ganti, et al., 1999; Burn-Thornton, et al., 1998; Ali and Wallace, 1997; Berson and Smith, 1997; Taha, et al., 1997; Fayyad, et al., 1996; Berson,

1996; Brachman and Anand, 1996; Adriaans and Zantinge, 1996; Chen, et al., 1996; Silberschatz and Tuzhilin, 1995; Piatetsky-Sharipo 1995; Canavos and Miller, 1995; Parsaye and Chignell 1993; Cercone and Tsuchiya 1993; Piatetsky-Sharipo 1992; Inmon and Osterfelt 1991; Everitt, 1980; Anderberg, 1973. The table summarizes the data mining techniques and illustrates why a certain technique has been chosen or not in this research.

Technique	Goal	Task	Justification
• ***Techniques that will be employed in this research:***			
SQL	Description	Summarization	-Easy to build and understand. -Preliminary analysis to reach the shallow knowledge. -Helpful in providing "general statistics" which most managers are interested in.
Visualization	Description	Summarization	-Easy to understand and requires no user background. -Can reach shallow and multi-dimensional knowledge. - Depends on the human side of the analysis.
Clustering analysis	Description Prediction	Clustering	-Achieves both data mining goals. -Works with many business applications. -Reaches hidden knowledge. -Able to handle a mixture of data types.
• ***Techniques that will NOT be employed in this research:***			
OLAP	-Needs special multi-dimensional DB formats. -Works best with complex mathematical functions. -Both SQL and Visualization can be used as alternatives to this technique.		
Association rules	-Result in useless rules unless we have an idea of the information we are looking for. -May result in missing some useful rules. -Most of the time it is not easy to determine the importance of the rules found. -May result in a huge number of rules that make it very difficult for a human being to analyze.		

ANN	-Needs extremely large training data sets. -Not easy to build, understand, or deploy. -Requires the user to understand the process. -Does not detect outliers' data. -Works best with complex problems such as fraud detection, biological simulations, and automated driving of unmanned vehicles.
GA	-Problem is complex and includes very large number of variables. -The GA parameters (population size, mutation rate, and number of generations) affect its performance. -Requires the user to understand the process. -Most of the time the GA output results need to be optimized through a linkage with an ANN.
Decision tree	-Handles raw data without preprocessing. -Easy to use and gives clear results. -Detects outliers. -Can handle text as well as numeric data. -Unable to handle multi-variable applications. -Can only handle one variable at a time. -Resources consuming (i.e. computer memory). -It takes a long time to build a decision tree algorithm.
Probabilistic graphical dependency	-Irrelevant, since the data used is not probabilistic.

Table (5-1). Data mining techniques' characteristics.

In the next sections the data mining techniques (SQL, Visualization, and Cluster analysis) that have been chosen to be used in this research will be evaluated in detail.

5-5 Standard Query Language (SQL)

The SQL statements will be used as a preliminary tool to find out shallow knowledge. It will also be used to summarize and describe the data set.

On the other hand, to respond to the DW users' needs (i.e. reports), SQL statements will be used to generate these requested reports (Refer to Appendix D for details). Details of these

SQL statements will be found in the ARDSS technical manual (Refer to Appendix F for details).

5-6 Visualization

The following benefits are earned beyond the use of visualization techniques (Berson, 1996):

- Users can easily interact with attributes and illustrate how they affect certain phenomenon;
- Users can view summarized data with drill down capability;
- Find multi-dimensional knowledge;
- It is considered an exploratory data analysis tool (EDA). That is data navigation, comparison, scaling, filtering are all available to users.

Visualization techniques will be used as a preliminary tool to represent the output of the SQL statements to the DSS users. The objectives of using visualization are:

1. Data summarization in the early stage of the analysis;
2. Finding out shallow knowledge;
3. Finding multi-dimensional knowledge;
4. Representing some reports which are based on the data stored in the data warehouse (Refer to Appendix D for details).

5-7 Clustering analysis

Clustering is basically classifying unclassified data (Larsen and Aone, 2000; Gordon, 1981; Everitt, 1980). The *data* to be classified consists of a set of *items* (sometimes referred to as *objects*, or *records*). Each item is described by a set of characteristics called *variables* (sometimes referred to as *attributes*). The target of clustering is to classify the items in the data set into a number of *groups* (sometimes referred to as *classes*, or *clusters*),

such that objects within one group have *similarities* to one another. Where the number of items to be grouped is n, the maximum number of groups should be n-1. In other words the number of expected groups varies from 2 to n-1.

The clustering process is done in two steps; *firstly* to find out the similarity or distance between each pair of items, *secondly* based on the similarity/distance matrix a clustering technique will be used to find out the number of groups and items within. Finally the groups found are depicted in a graph called *dendrogram*. There are 2^{n-2} ways of arranging a number of n cases in a dendrogram. When the analyst is not interested in all the clusters or groups the dendrogram could be *truncated* to represent the required number of clusters. For example, assume that 1000 student records are to be classified into 10 groups, then the dendrogram will be truncated to reflect those required ten groups and each of which will be described by a *cluster profile* (Wishart, 1999).

There are many methods to find the similarity or distance between items; these methods are called *proximity measures*. All proximity measures end up with either a *similarity matrix* (if it is a similarity measure) or a *distance matrix* (if it is a distance measure). S*i* denotes the similarity matrix, whilst D*i* denotes the distance matrix. Whether it is S_i or D_i the following rules hold true. Rule (1) $S_{ij} = S_{ji}$ (or alternatively $D_{ij} = D_{ji}$); that is to say that the matrix is *symmetric* (Wishart, 1998). Rule (2) $S_{ii} = 1$ (or $D_{ii} = 0$). Based on the last two rules, and since the similarity or distance matrix is *symmetric* this for storage purposes only $n\,(n\text{-}1)\,/\,2$ items need to be stored; that is either the lower or the upper triangular (Gordon, 1981).

Clustering analysis is able to achieve both the data mining goals, which are description and Prediction. By classifying the data into clusters or groups the description goal can be achieved. The prediction goal can be achieved by predicting which cluster a new record will join based on its attribute values and the nearest cluster to it (or the nearest profile in case of a dendrogram truncation took place).

In the following sections the different types of variables will be explained since they affect the choices of the clustering method, then difficulties in measuring variables will be explained, after that different proximity measures will be applied to a sample data set, and the data will then be clustered using various clustering techniques. Finally, analysis of the results will justify which proximity measure and clustering technique(s) that are to be used in this research.

5-7-1 Variable types and proximity measures

It is worth mentioning here that the outputs of clustering techniques-the clusters and entities- depend to a large extent on the input similarity or distance matrices between variables (Everitt, 1980). Similarity coefficients-sometimes referred to as *association coefficients*- measure the relationship between two individual items based on a number of variables. Similarity coefficients take values in the range from 0 to 1. Similarity measures are different from metric measures in many things foremost among which is their value. Whilst similarity measures usually take values between 0 and 1, distance metric measures take any positive value. The use of similarity or distance measures depends on the type of variables being studied (Everitt, 1980). There are actually two cases:

1. Variables are all of the same type. I.e. all variable are either binary (sometimes referred to as *Boolean*), or qualitative (sometimes referred to as *categorical*), or quantitative.
    a. *Binary data types* can be summarized in a two-way association table; many similarity coefficients have been proposed e.g. Gower[1], ROCK, Euclidean, Canberra, City Block, or Jaccard;
    b. For *quantitative data types* the product moment correlation coefficient is the most commonly used coefficient, Euclidean, Canberra, City Block or Gower could also be used. Although the correlation coefficient has been used as a measure of similarity, it has been criticized to the extent that some

[1] Gower and Jaccard have many formulas each fits with a specific data type.

scholars (Everitt, 1980; Wishart, 1971; Eades, 1965 in Everitt, 1980) said that it should not be used as a similarity measure any more. The correlation coefficient has been criticized on many backgrounds foremost among which is that it best fits with the linear data values (i.e. where there is a nonlinear perfect relationship the correlation coefficient is unable to describe it[2] properly);

c. For the *qualitative data types (Given that the data must be coded)* Gower, Euclidean, Canberra, City Block, or ROCK could also be used as similarity or distance measures (Guha, et al., 2000; Zhang, et al., 1996; Jain and Dubes, 1988; Gordon, 1981; Everitt, 1980; Anderberg, 1973; Gower, 1971).

2. Variables are of different types. I.e. Mixture of two or three variable types. In this case Gower is the most frequently used proximity measure (Guha, et al., 2000; Zhang, et al., 1996; Jain and Dubes, 1988; Gordon, 1981; Everitt, 1980; Anderberg, 1973; Gower, 1971).

5-7-2 Proximity measures difficulties

1. Missing values. In many classification data sets missing values is a common problem. Many approaches for handling missing data have been represented in the literature. However, since the data set under study in this thesis contains no missing values these approaches will not be discussed here. Missing values' techniques can be found in (Wishart, 1998; Gordon, 1981; Everitt, 1980; Anderberg, 1973).
2. Conditionally existing variables. This problem happens when a question is answered only if a certain condition occurs (i.e. if the condition does not occur the value does not exist). For example a discount is dependent on having a membership card, where the customer has the card the transaction is discounted, if not the discount value does not

[2] For example the three records (*R1*:2,4,6,8,10; *R2*:4,16,36,64,100; *R3*:1.2,3.3,5.1,7,8); where R2 is always the squared value of R1, whilst R1 and R3 do not obey a certain function all the time. However, the correlation coefficient between R1 and R3 is 0.994, and between R1 and R2 is 0.981. This is because the correlation is always high whenever the two records are linearly related.

show up. No such cases were found in the data sets under study in this thesis, so the techniques that handle this problem are also out of the research scope. Conditionally-exist handling techniques can be found in (Gordon, 1981; Everitt, 1980; Anderberg, 1973).

3. Incompatible variable units. This problem happens when the values in a certain variable are in different units. For example salaries are in US dollars for some employees, in GB Sterling for others, in Egyptian Pound for the rest. One simple solution for this problem is to transfer the values into one unit of measurement. More techniques are found in (Gordon, 1981). This problem was not found in the data set under study in this research.

Assume that a classification task will be undertaken to a group the employees whose salaries are reveived in different local currencies, it is very difficult to compare their salaries if not in one currency. The following table (5-2) provides an example.

Employee	**Salary (1000s)**	**Currency**	**Salary Class***	**Salary**	**Salary Class****	**Class changed?**
	Before transfer			***After transfer***		
E.1	8	GBP	Class E	11.2	Class D	Y
E.2	12	USD	Class D	12	Class D	N
E.3	25	L.E	Class C	6.25	Class E	Y
E.4	45	L.E	Class B	11.25	Class D	Y

* <= 10 Class E; 10 < Class D <=20; 20 < Class C < 30; 30 < Class B <=40; 40 < Class A
** GBP =1.4 USD, USD= 4 L.E.

Table (5-2). The transfer effect.

4. Weighting of variables. Some of the pre-stated proximity measures have two forms; one used when the variables are equally weighted; another is used when variables own different weighting factors. Examples are Gower and Euclidean. Based on the problems associated with variable weighting some scholars advocate that each variable should be weighted equally (Gordon, 1981; Sneath and Sokal, 1973). The argument being that if the variables are not equally weighted there will be difficulties in

justifying a specific value to a certain variable. Some scholars do not pay much attention to the weighting problems (Guha, et al., 2000; Everitt, 1980; Anderberg, 1973). Based on this illustration, the variables under study in this research are equally weighted.

5. Variables of mixed data types. It is common also in classification data sets that variables under study are of mixed data types. The data under study in this research is also of mixed types (i.e. Gender is qualitative coded binary, major is categorical, Grade is quantitative.... Etc.). There are three main approaches that handle this problem:
    a. Convert all the variables to the most frequently data type in the data set. However, this transformation normally discards valuable information (Gordon, 1981; Anderberg, 1973);
    b. Carry out separate classification analysis for each group of variables. This is the least recommended approach (Gordon, 1981);
    c. Employ a general similarity coefficient that can deal with mixture of variable types (Gordon, 1981; Everitt, 1980). An appropriate coefficient is Gower (1971).

Based on this evidence the approach used here will be to choose a similarity/distance measure that is appropriate for handling mixed data types.

5-7-3 Clustering algorithms[3]

To choose the most suitable clustering algorithm for the students' data sets under study by this thesis, the following steps have been undertaken:

1. A sample of seven records was drawn from the 2000 record data set. The sample of seven records is then classified into three main subgroups. The first subgroup consists of two records that are almost typical-*similar*, the second subgroup consists of two

[3] An algorithm is any well-defined computational procedure that takes set of values as input and set of values as output, an algorithm is thus a sequence of computational steps to transform the input into output (Cormen, et al., 2000:1).

records that are almost different-*dissimilar*, and the last subgroup consists of four records that are between similarity and dissimilarity;

2. The seven records sample drawn from the main 2000 records sample is a purposive sample for the reason of evaluating the results of different algorithms applied to:
    2.1 Similar data records;
    2.2 Dissimilar data records;
    2.3 Fairly similar or dissimilar records;
3. To choose the most suitable clustering algorithm, two steps are to be fulfilled:
    3.1 Choosing the most suitable proximity measure;
    3.2 Choosing the most suitable clustering technique;
4. Different proximity measures have been applied to the sample of seven records in order to choose the one that gives the most reasonable results. For example, if two records are similar but the proximity measure indicates that they are not, this would be a reason to eliminate this measure or to say that the measure is irrelevant to this type of data;
5. Different clustering techniques have been applied to the sample record set given certain proximity measures;
6. The pair of both proximity measure and clustering technique that will be chosen here will be used as the clustering algorithm to run on the 2000 records set.

5-7-4 The sample record set

5-7-4-1 Why seven records?

The decision to choose only seven records to decide on the appropriate proximity measure and clustering technique is justified as follows:

1. The previous research work done to choose the appropriate proximity measure and clustering technique for a certain data set:

a. Everitt (1980: 17) used 3 records of 10 variables to describe the proximity measures. He also used only 5 records to decide on any technique's appropriateness;
b. Gordon (1981: 35) used 7 records of 2 variables decide on any technique's appropriateness;
c. Guha, et al. (2000: 347) used 4 records of 6 attributes for the same purpose;

2. The seven records are representative in the sense that they include all the eleven attributes that are used to describe the 2000 students, and they combine different cases of similarity (i.e. similar, dissimilar, and fairly similar records);
3. There is no intention to derive any knowledge out of these seven records about the entire sample.

5-7-4-2 The sample

The following table (5-3) includes the sample record set:

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11
Student 1	1	11	92	92	65	8	98	1	74	11	1
Student 2	1	11	92	92	66	8	98	1	74	11	1
Student 3	1	11	92	92	75	8	97	6	75	11	0
Student 4	1	0	89	89	76	10	95	2	71	1	1
Student 5	1	0	92	92	54	2	97	6	90	11	0
Student 6	2	11	92	92	64	2	96	6	75	11	0
Student 7	1	0	91	92	91	9	98	1	74	11	1

Table (5-3). Sample record set.

Where **Var 1** is the high school code, **Var 2** is the high school country, **Var 3** is the high school year, **Var 4** is the year join the university, **Var 5** is the high school percent, **Var 6** is the major, **Var 7** is the graduation year, **Var 8** is the graduation grade, **Var 9** is the date of birth, **Var 10** is the nationality and **Var 11** is the Gender. For the codes of these data records please refer to Appendix (C).

Subgroup 1: Similar data records

From the sample we can notice that records (**Student 1** and **Student 2**) are almost identical or similar. They have same values for ten variables, while only one variable is different.

Subgroup 2: Dissimilar data records

Records (**Student 4** and **Student 6**) are almost different or dissimilar, all the values for the eleven variables are different in both records.

Subgroup 3: Fairly similar or dissimilar records

Records (**Student 3** and **Student 5**) are fairly similar. They have same values for seven variables and different four values for the remaining four variables. Records (**Student 5** and **Student 7**) are fairly dissimilar. They have same values for four variables and different seven values for the remaining seven variables.

5-7-4-3 The sample characteristics

Some of the proximity measures require the Standard deviation (**Modified Euclidean metric**) of the data and others require the range (**Gower similarity measure**).

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11
StDev	0.38	5.88	1.13	1.13	11.8	3.3	1.15	2.56	6.26	3.78	0.53
Range	1	11	3	3	37	8	3	5	19	10	1

Table (5-4). Sample characteristics.

5-7-5 Which proximity measure to use in this research?

Based on the previous related work done by (Guha, et al., 2000; Zhang, et al., 1996; Jain and Dubes, 1988; Gordon, 1981; Everitt, 1980; Anderberg, 1973; Gower, 1971), Gower is the most relevant proximity measure to the data set under study in this research. However, to examine the consistency of the results, different proximity measures will be applied to the data set. The measure that will give the most consistent results will then be employed.

5-7-5-1 Analyzing various proximity measures against the different groups of records

The next sections are based on applying the different proximity measures to the record sets, and then evaluates the results before deciding which measure will be used.

5-7-5-2 The various proximity measures to be analyzed by this study

The following table (5-5) summarizes the various proximity measures that will be used to calculate either the similarities or distances between the different groups of records under study.

Proximity measure	Type	Cited in literature	Formula
Jaccard	**Similarity**	(Guha, at al., 2000: 347; Jain and Dubes, 1988: 17; Gordon, 1981: 19; Everitt, 1980: 13; Anderberg, 1973: 117)	$S_{ij} = (T_1 \cap T_2) / (T_1 \cup T_2)$ Where T1 and T2 are data records
Gower	**Similarity**	(Gower, 1971 in Jain and Dubes, 1988: 17; Gordon, 1981: 23; Everitt, 1980: 16; Anderberg, 1973: 141)	$S_{ij} = 1 - \lvert X_{ik} - X_{jk} \rvert / R_k$ Where R is the range of variable k
Euclidean	**Distance**	(Guha, et al., 2000: 346; Wishart, 1998: 260; Zhang, et al., 1996: 103; Jain and Dubes, 1988: 15; Gordon, 1981: 21; Everitt, 1980: 17; Anderberg, 1973: 100)	$D_{ij} = \{ \sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \}^{1/2}$
Modified Euclidean	**Distance**	(Everitt, 1980: 20)	$Z_{ij} = X_{ik} / \sigma$ Where σ is the standard deviation of the k_{th} variable
City Block	**Distance**	(Gordon, 1981: 21)	$D_{ij} = \sum_{k=1}^{n} \lvert X_{ik} - X_{jk} \rvert$
Canberra	**Distance**	(Lance and Williams, 1966 in Gordon, 1981: 21; Anderberg, 1973: 112)	$d_{ij} = \sum_{k=1}^{p} [\lvert X_{ik} - X_{jk} \rvert / (X_{ik} + X_{jk})]$

Table (5-5). Summary of the proximity measures.

- Note that the City Block metric and Euclidean metric proximity measures are the special cases $\lambda = 1$ and $\lambda = 2$ of the **Minkowski** metrics (in Jain and Dubes, 1988: 14; Gordon, 1981: 21; Everitt, 1980: 19; Anderberg, 1973: 101):

$$d_{ij}^{\lambda} = \{ \sum_{k=1}^{p} |Xik - Xjk|^{\lambda} \}^{1/\lambda} \qquad \text{where } \lambda > 0$$

5-7-6 The different groups of records

In the next sections the three groups of records will be illustrated.

5-7-6-1 The first group: similar records

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11
Student 1	1	11	92	92	65	8	98	1	74	11	1
Student 2	1	11	92	92	66	8	98	1	74	11	1
Difference	**0**	**0**	**0**	**0**	**-1**	**0**	**0**	**0**	**0**	**0**	**0**
Similarity count	**10**										
Dis-similarities	**1**										
Count all	**11**										

Table (5-6). Similar records; student 1 and 2.

5-7-6-2 The second group: dissimilar records

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var10	Var11
Student 4	1	0	89	89	76	10	95	2	71	1	1
Student 6	2	11	92	92	64	2	96	6	75	11	0
Difference	**-1**	**-11**	**-3**	**-3**	**12**	**8**	**-1**	**-4**	**-4**	**-10**	**1**
Similarity count	**0**										
Dis-similarities	**11**										
Count all	**11**										

Table (5-7). Dissimilar records; student 4 and 6.

5-7-6-3 The third group: fairly similar records

5-7-6-3-1 Similar in seven, different in four variables

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var10	Var11
Student 3	1	11	92	92	75	8	97	6	75	11	0
Student 5	1	0	92	92	54	2	97	6	90	11	0
Difference	**0**	**11**	**0**	**0**	**21**	**6**	**0**	**0**	**-15**	**0**	**0**
Similarity count	**7**										
Dis-similarities	**4**										
Count all	**11**										

Table (5-8). Fairly similar records; student 3 and 5.

5-7-6-3-2 Similar in four, different in seven variables (opposite of the last case)

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var10	Var11
Student 5	1	0	92	92	54	2	97	6	90	11	0
Student 7	1	0	91	92	91	9	98	1	74	11	1
Difference	**0**	**0**	**1**	**0**	**-37**	**-7**	**-1**	**5**	**16**	**0**	**-1**
Similarity count	**4**										
Dis-similarities	**7**										
Count all	**11**										

Table (5-9). Fairly similar records; student 5 and 7.

5-7-7 Applying the various proximity measures to the different groups of records

The results of applying the proximity measures to the sample data set (i.e. groups of records) are presented in Appendix (E). The next section evaluates these results.

5-7-8 Evaluating the proximity measures

The following table (5-10) shows the performance of all the pre-defined proximity measures on the three different groups of records.

Proximity Measures	Record Groups			
	1st Similar	2nd Dissimilar	3rd Fairly Similar	
			1st Fairly	2nd Fairly
Jaccard	.83	0	.47	.22
Gower	.99	.21	.72	.51
Euclidean	1	21.95	28.69	41.3
Euclidean Sim. Equivalent	0.5	0.04	0.03	.02
Modified Euclidean	.08	6.7	3.96	5.47
Modified Euclidean Sim.Equivalent	.92	.13	.20	.15
City Block metric	1	58	53	68
City Block metric Equivalent	.5	.017	.02	.01
Canberra metric	.01	4.47	1.85	2.68
Canberra metric Equivalent	.99	.18	.35	.27

Table (5-10). Proximity measures comparison.

To be able to compare the results of the proximity measures' to each other, we need to transform the metric measures to their similarity measure equivalent values. Euclidean, Modified Euclidean, City Block, and Canberra metric measures have been transformed into an equivalent similarity measure value using the following formula (Gordeon, 1981: 13; Everitt, 1980: 19):

$$S_{ij} = 1 / (1 + d_{ij})$$

5-7-9 Discussion on "*Which proximity measure to use for the data set?*"

1. The research results are consistent with the previous related work in two findings:
    a. Previous related work as discussed earlier indicates that **Gower** is the most suitable measure where the data types are mixture. Gower similarity coefficient scores are: 0.99 for the almost similar records, 0.21 for the dis-similar records, 0.72 for the fairly similar, and 0.51 for the fairly dis-similar.

These scores are good in the sense that Gower was able to detect similarities (0.99), dissimilarities (0.21), and for the fairly similar records (3rd group) the measure was able to differentiate between two groups on of which has of 7 attributes in common (0.72) and the other of four attributes in common (0.51);

b. Previous work also did not recommend **Jaccard** where the data types are mixture. For the data under study, Jaccard is the only measure that produced an extreme value for any of the groups (i.e. 0 or 1). Jaccard is based on the number of similarities and dissimilarities and it completely discards the attribute values, it is possible that its value could be 1 or 0. For the students' data, if one student has got 90 % and the other 85 % this does not mean that the similarity between them is zero; there is still a degree of similarity between the two. However, since Jaccard is based on counting the similarity and dissimilarities, the score was 0.0 for the second group and it is the only measure that did not detect any value for this group. Jaccard produced 0.83 for the first group which is better than both Euclidean and City block for the same group, but lower than the values of Gower and modified Euclidean;

2. The **City block** coefficient will not be chosen because of the following reasons:
    a. The City block coefficient produced a value of 0.5 for the first group that includes two very similar records. That makes it incomparable with the 0.99 value produced by Gower or the 0.92 produced by modified Euclidean for the same group;
    b. Also according to this coefficient, the similarity of the records in the second (i.e. dissimilar records) .017 is higher than the similarity of the records in the last group .01 (fairly dissimilar records);

3. Although **Euclidean** metric measure did not perform consistently well as Gower and modified Euclidean, it will be retained and used with the clustering techniques because of the following reasons:

    a. Euclidean was recommended to be used with many clustering techniques (e.g. Nearest Neighbour, Furthest Neighbour, Ward's) by many scholars (Wishart, 2000; Gordon, 1981; Everitt, 1980);

    b. The results of the Euclidean measure are good because it was able to distinguish between similar and dissimilar records;

4. Another finding of this research is the results obtained by the **modified Euclidean** and **Canberra metric** measures (after transforming their results from distance metric to equivalent similarity coefficients). Previous work shows that little analysis has been carried out to apply the modified Euclidean and Canberra metric measures to a mixture of data types. However, the results of this research indicate that they are in close agreement with the results obtained by Gower's similarity measure. The following table (5-11) summarizes this finding:

Proximity Measures	Record Groups			
	1st Similar	2nd Dissimilar	3rd Fairly Similar	
			1st Fairly	2nd Fairly
Gower	.99	.21	.72	.51
Euclidean Sim. Equivalent	.5	.04	.03	.02
Modified Euclidean Sim. Equivalent	.92	.13	.20	.15
Canberra metric Equivalent	.99	.18	.35	.27

Table (5-11). Relevant proximity measures.

5. The decision on which proximity measure to use will not be taken now. This analysis indicates that these proximity measures may be appropriate (i.e. Gower, Euclidean, modified Euclidean, and Canberra metric) to be use with a mixture of data types. But before deciding which measure is "best", the performance of these proximity measures will be evaluated when using different clustering techniques.

5-7-10 The proximity matrix for the sample data set

In the next sections *Gower*, *Euclidean*, *Modified Euclidean*, and *Canberra* proximity measures' matrices will be prepared as they are required to be used by the clustering techniques.

5-7-10-1 Gower similarity matrix

$$\mathbf{S_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ .998 & 1 & & & & & \\ .759 & .761 & 1 & & & & \\ .463 & .466 & .368 & 1 & & & \\ .525 & .523 & .718 & .267 & 1 & & \\ .591 & .589 & .784 & .212 & .692 & 1 & \\ .804 & .806 & .611 & .586 & .511 & .395 & 1 \end{pmatrix} \end{array}$$

5-7-10-2 Euclidean distance matrix

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ 1 & 0 & & & & & \\ 11.3 & 10.4 & 0 & & & & \\ 19.6 & 19 & 16.8 & 0 & & & \\ 23.7 & 24.2 & 28.7 & 32.4 & 0 & & \\ 8.3 & 8.5 & 12.6 & 22 & 21.2 & 0 & \\ 28.3 & 27.3 & 20.2 & 19 & 41.2 & 30.5 & 0 \end{pmatrix} \end{array}$$

5-7-10-3 Modified Euclidean distance matrix

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ .085 & 0 & & & & & \\ 2.98 & 2.96 & 0 & & & & \\ 5.74 & 5.73 & 5.86 & 0 & & & \\ 4.7 & 4.74 & 3.96 & 6.97 & 0 & & \\ 4.5 & 4.54 & 3.44 & 6.75 & 4.22 & 0 & \\ 3.04 & 2.98 & 3.79 & 5.1 & 5.47 & 5.6 & 0 \end{pmatrix} \end{array}$$

5-7-10-4 The Canberra distance matrix

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ .008 & 0 & & & & & \\ 1.798 & 1.790 & 0 & & & & \\ 2.425 & 2.418 & 3.522 & 0 & & & \\ 3.509 & 3.517 & 1.854 & 2.331 & 0 & & \\ 2.672 & 2.680 & 1.018 & 4.485 & 1.514 & 0 & \\ 1.231 & 1.224 & 2.887 & 1.373 & 2.714 & 3.881 & 0 \end{pmatrix} \end{array}$$

5-7-11 Evaluating the clustering techniques

There are a number of clustering techniques that are used with different types of data in different occasions, based on different proximity measures. According to (Wishart, 1998; Gordon, 1981; Everitt, 1980; Anderberg, 1973) these techniques are the following:

1. Hierarchical techniques (agglomerative and divisive);
2. Optimization techniques;
3. Density search techniques;
4. Clumping techniques.

In the following section these techniques will be introduced and some will be applied to the sample data set. Then, an evaluation of the results of these techniques will identify the proximity measure(s) and clustering technique(s) that will be employed by this thesis.

5-7-11-1 Hierarchical Clustering techniques: I- Agglomerative techniques

1- Nearest neighbour (NN) (in Wishart, 1998: 261; Jain and Dubes, 1988: 60; Gordon, 1981: 35; Everitt, 1980: 25; Anderberg, 1973: 137)

This clustering technique can be used with either similarity or distance proximity measures. After formulating the proximity matrix the smallest item is found (distance matrix) OR the largest item is found (similarity matrix) and the two items are then fused together to formulate a new item, each fusion decreases the number of items by one. The distance between groups is defined as the distance between their closest members. This NN technique ends with a diagram called *single linkage dendrogram* which shows the group fusions.

-The next example shows the results of applying NN to the sample data records which similarities have been calculated using Gower coefficient:

$$\mathbf{S_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ .998 & 1 & & & & & \\ .759 & .761 & 1 & & & & \\ .463 & .466 & .368 & 1 & & & \\ .525 & .523 & .718 & .267 & 1 & & \\ .591 & .589 & .784 & .212 & .692 & 1 & \\ .804 & .806 & .611 & .586 & .511 & .395 & 1 \end{pmatrix} \end{array}$$

The largest entry in $\mathbf{S_1}$ (.998) is S_{12}. Therefore items 1 and 2 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(12)3} = \text{MAX}\{S_{13}, S_{23}\} = S_{23} = .761$$

$$S_{(12)4} = \text{MAX}\{S_{14}, S_{24}\} = S_{24} = .466$$

$$S_{(12)5} = \text{MAX}\{S_{15}, S_{25}\} = S_{15} = .525$$

$$S_{(12)6} = \text{MAX}\{S_{16}, S_{26}\} = S_{16} = .591$$

$$S_{(12)7} = \text{MAX}\{S_{17}, S_{27}\} = S_{27} = .806$$

The new matrix is now $\mathbf{S_2}$:

$$\mathbf{S_2} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccc} 12 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 1 & & & & & \\ .761 & 1 & & & & \\ .466 & .368 & 1 & & & \\ .525 & .718 & .267 & 1 & & \\ .591 & .784 & .212 & .692 & 1 & \\ .806 & .611 & .586 & .511 & .395 & 1 \end{bmatrix} \end{array}$$

The largest entry in $\mathbf{S_2}$ (.806) is $S_{(12)\,7}$. Therefore items 12 and 7 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(12\,7)\,3} = \text{MAX}\,\{S_{13}, S_{23}, S_{73}\} = S_{23} = .761$$

$$S_{(12\,7)\,4} = \text{MAX}\,\{S_{14}, S_{24}, S_{74}\} = S_{74} = .586$$

$$S_{(12\,7)\,5} = \text{MAX}\,\{S_{15}, S_{25}, S_{75}\} = S_{15} = .525$$

$$S_{(12\,7)\,6} = \text{MAX}\,\{S_{16}, S_{26}, S_{76}\} = S_{16} = .591$$

The new matrix is now $\mathbf{S_3}$:

$$\mathbf{S_3} = \begin{array}{c} \\ 127 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 127 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1 & & & & \\ .761 & 1 & & & \\ .586 & .368 & 1 & & \\ .525 & .718 & .267 & 1 & \\ .591 & .784 & .212 & .692 & 1 \end{bmatrix} \end{array}$$

The largest entry in $\mathbf{S_3}$ (.784) is $S_{(63)}$. Therefore items 6 and 3 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(63)\,127} = \text{MAX}\,\{S_{61}, S_{62}, S_{67}, S_{31}, S_{32}, S_{37}\} = S_{32} = .761$$

$$S_{(63)\,4} = \text{MAX}\,\{S_{64}, S_{34}\} = S_{34} = .368$$

$$S_{(63)\,5} = \text{MAX}\,\{S_{65}, S_{35}\} = S_{35} = .718$$

The new matrix is now $\mathbf{S_4}$:

$$\mathbf{S_4} = \begin{array}{c} \\ 127 \\ 36 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 127 & 36 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & & \\ .761 & 1 & & \\ .586 & .368 & 1 & \\ .525 & .718 & .267 & 1 \end{bmatrix} \end{array}$$

The largest entry in $\mathbf{S_4}$ (.761) is $S_{7(127)(36)}$. Therefore items 127 and 36 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(127\,36)\,4} = \text{MAX}\{S_{14}, S_{24}, S_{74}, S_{34}, S_{64}\} = S_{74} = .586$$

$$S_{(127\,36)\,5} = \text{MAX}\{S_{15}, S_{25}, S_{75}, S_{35}, S_{65}\} = S_{35} = .718$$

The new matrix is now $\mathbf{S_5}$:

$$\mathbf{S_5} = \begin{array}{c} \\ 12736 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 12736 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & \\ .586 & 1 & \\ .718 & .267 & 1 \end{bmatrix} \end{array}$$

The largest entry in $\mathbf{S_5}$ (.718) is $S_{(12736)\,5}$. Therefore items 12736 and 5 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(12736\,5)\,4} = \text{MAX}\{S_{14}, S_{24}, S_{74}, S_{34}, S_{64}, S_{54}\} = S_{74} = .586$$

The new matrix is now $\mathbf{S_6}$:

$$\mathbf{S_6} = \begin{array}{c} \\ 127365 \\ 4 \end{array} \begin{array}{c} \begin{array}{cc} 127365 & 4 \end{array} \\ \begin{bmatrix} 1 & \\ .586 & 1 \end{bmatrix} \end{array}$$

Finally fusion of the remaining two groups is done to formulate a one single group that contains all the 7 student records.

- *Interpretation*:

Group 1 (items 1 and 2) $0.998 >= S_{ij} > 0.806$

Group 2 (items 1, 2 and 7) $0.806 >= S_{ij} > 0.784$

Group 3 (items 3 and 6) $0.784 >= S_{ij} > 0.761$

Group 4 (items 1, 2, 7 and 3, 6) $0.761 >= S_{ij} > 0.718$

Group 5 (items 1, 2, 7, 3, 6 and 5) $0.718 >= S_{ij} > 0.586$

Group 6 (all items) $S_{ij} <= 0.586$

Sneath (1957), (in Anderberg, 1973: 41) introduced a single link algorithm that is very simple and is done in one step from the proximity matrix. Depending on S_1:

1 and 2 join at level 0.998

7 join (1 and 2) at level 0.806

3 join 6 at level 0.784

(3 and 6) join (1, 2, and 7) at level 0.761

5 join (1, 2, 7, 3, and 6) at level 0.718

4 join (1, 2, 7, 3, 6, and 5) at level 0.586

Sneath's algorithm is much simpler than the one presented by Everitt and Gordon.

-In the following example NN is applied to the data sets whose proximity is based on the modified Euclidean measure:

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ .085 & 0 & & & & & \\ 2.98 & 2.96 & 0 & & & & \\ 5.74 & 5.73 & 5.86 & 0 & & & \\ 4.7 & 4.74 & 3.96 & 6.97 & 0 & & \\ 4.5 & 4.54 & 3.44 & 6.75 & 4.22 & 0 & \\ 3.04 & 2.98 & 3.79 & 5.1 & 5.47 & 5.6 & 0 \end{pmatrix} \end{array}$$

Applying the NN to the items in $\mathbf{D}_1$ will result in the following groups:

1 and 2 join at level 0.085

3 join (1 and 2) at level 2.96

7 join (1, 2, and 3) at level 2.98

6 join (1, 2, 3, and 7) at level 3.44

5 join (1, 2, 3, 7, and 6) at level 4.22

4 join (1, 2, 3, 5, 7, and 6) at level 5.73

-In the following example NN is applied to the data sets whose proximity is based on the Euclidean measure:

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ 1 & 0 & & & & & \\ 11.3 & 10.4 & 0 & & & & \\ 19.6 & 19 & 16.8 & 0 & & & \\ 23.7 & 24.2 & 28.7 & 32.4 & 0 & & \\ 8.3 & 8.5 & 12.6 & 22 & 21.2 & 0 & \\ 28.3 & 27.3 & 20.2 & 19 & 41.2 & 30.5 & 0 \end{pmatrix} \end{array}$$

1 and 2 join at level 1

6 join (1 and 2) at level 8.3

3 join (1, 2, and 6) at level 10.4

4 join (1, 2, 6, and 3) at level 6.8

7 join (1, 2, 6, 3, and 4) at level 19

5 join (1, 2, 6, 3, 4, and 7) at level 23.7

-In the following example NN is applied to the data sets whose proximity is based on the Canberra metric measure:

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ .008 & 0 & & & & & \\ 1.798 & 1.790 & 0 & & & & \\ 2.425 & 2.418 & 3.522 & 0 & & & \\ 3.509 & 3.517 & 1.854 & 2.331 & 0 & & \\ 2.672 & 2.680 & 1.018 & 4.485 & 1.514 & 0 & \\ 1.231 & 1.224 & 2.887 & 1.373 & 2.714 & 3.881 & 0 \end{pmatrix} \end{array}$$

1 and 2 join at level 0.008

3 and 6 join at level 1.018

7 join (1 and 2) at level 1.224

4 join (1, 2 and 7) at level 1.373

5 join (3 and 6) at level 1.514

- Summary of NN:

The results of applying the NN to the different proximity measures are depicted in the following Dendrogram figures (5-1: 5-4). Section (5-7-12) analyzes these results in detail.

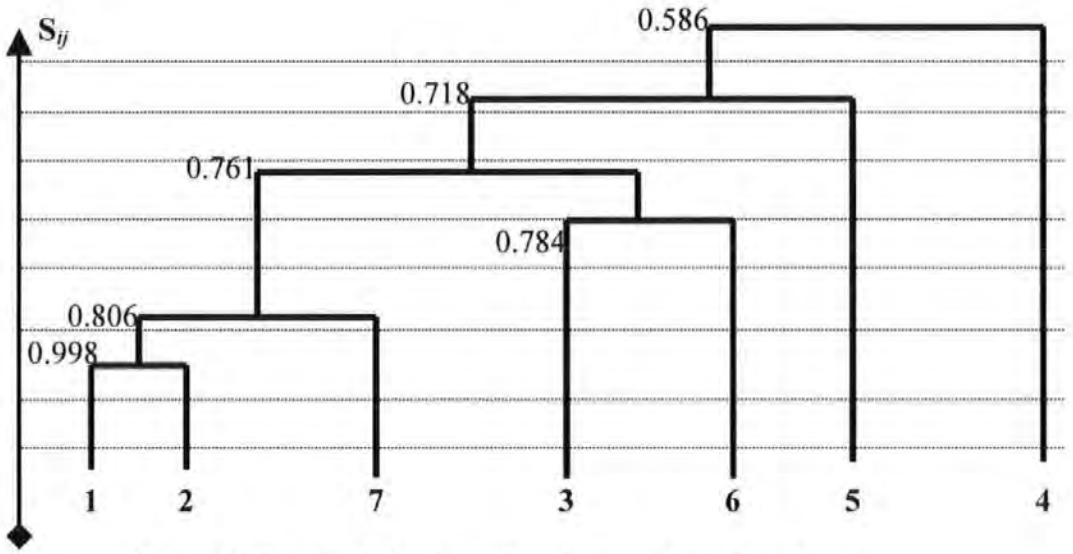


Figure (5-1). Single Linkage Dendrogram based on Gower.

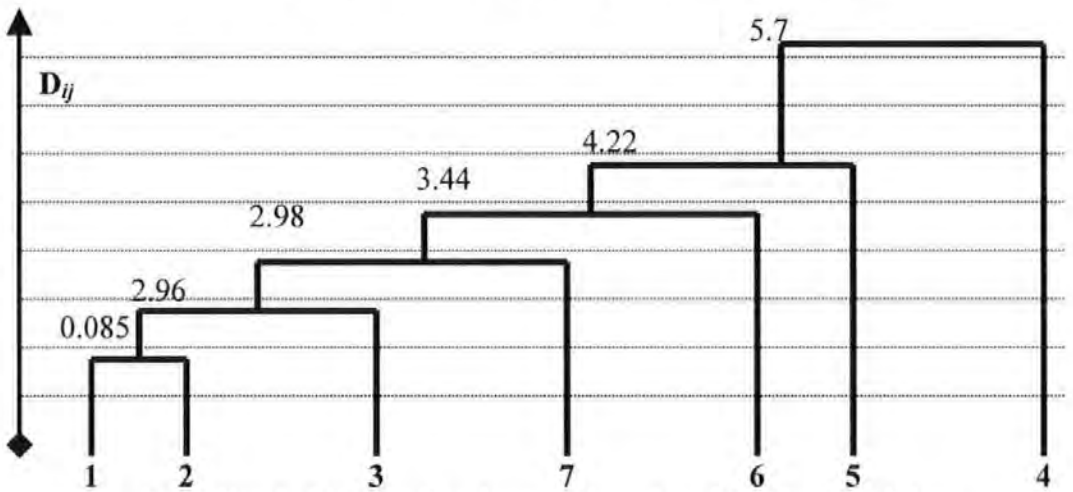


Figure (5-2). Single Linkage Dendrogram based on Modified Euclidean.

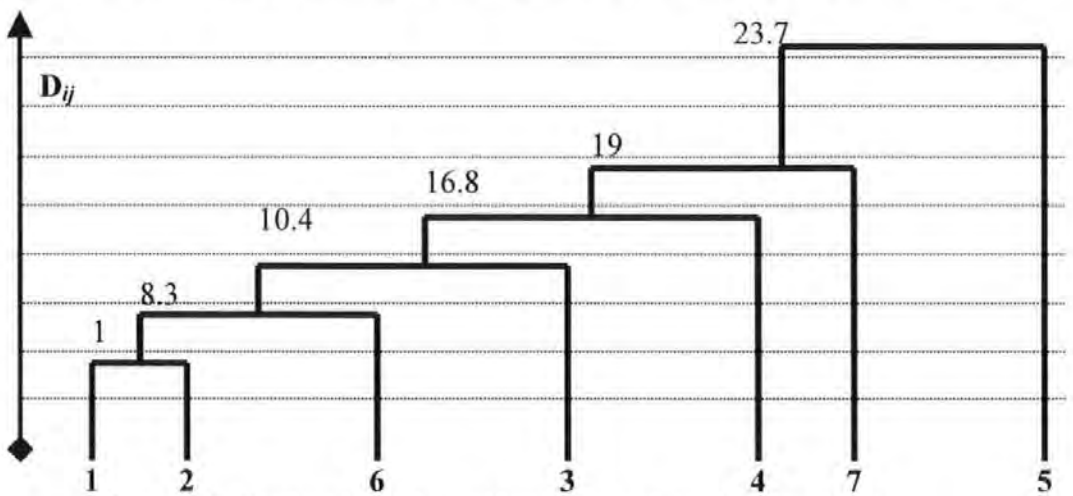


Figure (5-3). Single Linkage Dendrogram based on Euclidean.

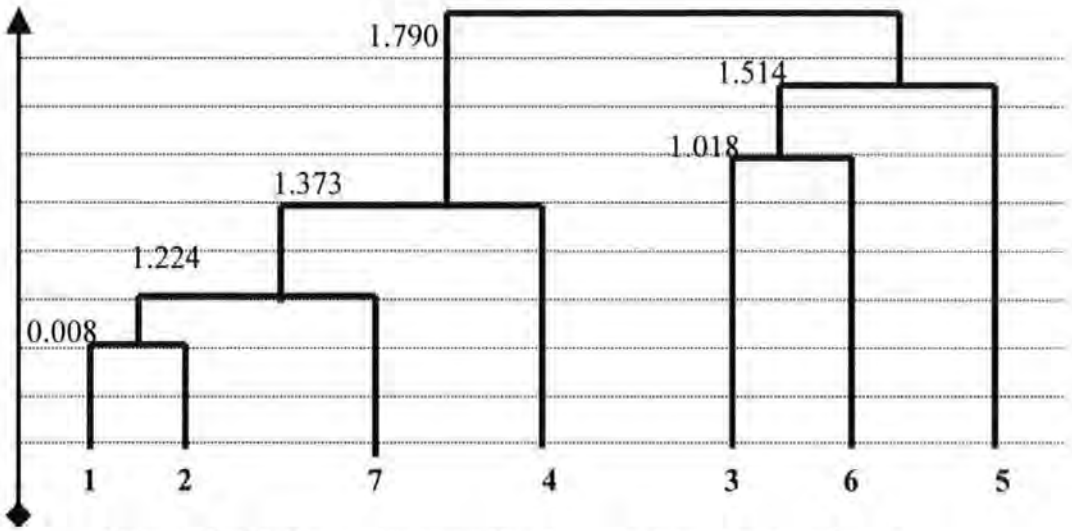


Figure (5-4). Single Linkage Dendrogram based on Canberra.

2- Furthest neighbour (FN) (in Jain and Dubes, 1988: 63; Everitt, 1980: 28; Anderberg, 1973: 138)

This technique starts by selecting the closest individuals to fuse or join together. In the case of similarity matrix this is represented by the largest value, whilst in the case of a distance matrix this is represented by the smallest value.

NN and FN are opposite in how they update the proximity matrix:

- NN uses the closest i.e. largest value if similarity measure is used, OR smallest value if distance measure is used;
- FN uses the furthest i.e. smallest value if similarity measure is used, OR largest value if distance measure is used.

FN ends with a diagram shows group fusions called *complete linkage dendrogram*. This technique falls into the agglomerative hierarchical techniques (like NN), uses the idea of group fusions. However, it does not always produce the same results as NN for the same data sets (Everitt, 1980).

-The next example shows the results of applying FN to the sample data records which similarities have been calculated using Gower coefficient:

$$\mathbf{S_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ .998 & 1 & & & & & \\ .759 & .761 & 1 & & & & \\ .463 & .466 & .368 & 1 & & & \\ .525 & .523 & .718 & .267 & 1 & & \\ .591 & .589 & .784 & .212 & .692 & 1 & \\ .804 & .806 & .611 & .586 & .511 & .395 & 1 \end{pmatrix} \end{array}$$

The closest entry in $\mathbf{S_1}$ (.998) is S_{12}. Therefore items 1 and 2 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(12)3} = \text{MIN}\{S_{13}, S_{23}\} = S_{13} = .759$$

$$S_{(12)4} = \text{MIN}\{S_{14}, S_{24}\} = S_{14} = .463$$

$$S_{(12)5} = \text{MIN}\{S_{15}, S_{25}\} = S_{25} = .523$$

$$S_{(12)6} = \text{MIN}\{S_{16}, S_{26}\} = S_{26} = .589$$

$$S_{(12)7} = \text{MIN}\{S_{17}, S_{27}\} = S_{17} = .804$$

The new matrix is now $\mathbf{S_2}$:

$$\mathbf{S_2} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccc} 12 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & & & & & \\ .759 & 1 & & & & \\ .463 & .368 & 1 & & & \\ .523 & .718 & .267 & 1 & & \\ .589 & .784 & .212 & .692 & 1 & \\ .804 & .611 & .586 & .511 & .395 & 1 \end{pmatrix} \end{array}$$

The closest entry in $\mathbf{S_2}$ (.804) is $S_{(12)7}$. Therefore items 12 and 7 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(12\,7)3} = \text{MIN}\{S_{13}, S_{23}, S_{73}\} = S_{73} = .611$$

$$S_{(12\,7)4} = \text{MIN}\{S_{14}, S_{24}, S_{74}\} = S_{14} = .463$$

$$S_{(12\,7)5} = \text{MIN}\{S_{15}, S_{25}, S_{75}\} = S_{75} = .511$$

$$S_{(12\,7)6} = \text{MIN}\{S_{16}, S_{26}, S_{76}\} = S_{76} = .395$$

The new matrix is now $\mathbf{S_3}$:

$$\mathbf{S_3} = \begin{array}{c} \\ 127 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 127 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1 & & & & \\ .611 & 1 & & & \\ .463 & .368 & 1 & & \\ .511 & .718 & .267 & 1 & \\ .395 & .784 & .212 & .692 & 1 \end{bmatrix} \end{array}$$

The closest entry in $\mathbf{S_3}$ (.784) is S_{36}. Therefore items 3 and 6 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(36)\,127} = \text{MIN}\,\{S_{31}, S_{32}, S_{37}, S_{61}, S_{62}, S_{67}\} = S_{67} = .395$$

$$S_{(36)\,4} = \text{MIN}\,\{S_{34}, S_{64}\} = S_{64} = .212$$

$$S_{(36)\,5} = \text{MIN}\,\{S_{35}, S_{65}\} = S_{65} = .692$$

The new matrix is now $\mathbf{S_4}$:

$$\mathbf{S_4} = \begin{array}{c} \\ 127 \\ 36 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 127 & 36 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & & \\ .395 & 1 & & \\ .463 & .212 & 1 & \\ .511 & .692 & .267 & 1 \end{bmatrix} \end{array}$$

The closest entry in $\mathbf{S_4}$ (.692) is $S_{(36)\,5}$. Therefore items 36 and 5 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(36\,5)\,127} = \text{MIN}\,\{S_{31}, S_{61}, S_{51}, S_{32}, S_{62}, S_{52}, S_{37}, S_{67}, S_{57}\} = S_{67} = .395$$

$$S_{(35\,6)\,4} = \text{MIN}\,\{S_{34}, S_{54}, S_{64}\} = S_{64} = .212$$

The new matrix is now $\mathbf{S_5}$:

$$\mathbf{S_5} = \begin{array}{c} \\ 127 \\ 365 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} 127 & 365 & 4 \end{array} \\ \begin{bmatrix} 1 & & \\ .395 & 1 & \\ .463 & .212 & 1 \end{bmatrix} \end{array}$$

The closest entry in $\mathbf{S_5}$ (.463) is $S_{(127)\,4}$. Therefore items 127 and 4 are fused together in one item, and the similarity between the new item and the rest is found as follows:

$$S_{(127\,4)\,365} = \text{MIN}\,\{S_{13}, S_{23}, S_{73}, S_{43}, \ldots, S_{45}\} = S_{46} = .212$$

The new matrix is now $\mathbf{S_6}$:

$$S_6 = \begin{array}{c} \\ 1274 \\ 365 \end{array} \begin{array}{c} \begin{array}{cc} 1274 & 365 \end{array} \\ \begin{bmatrix} 1 & \\ .212 & 1 \end{bmatrix} \end{array}$$

Finally fusion of the remaining two groups is done to formulate a one single group that contains all the 7 student records.

- *Interpretation*:

Group 1 (items 1 and 2) $0.998 >= S_{ij} > 0.804$

Group 2 (items 1, 2 and 7) $0.804 >= S_{ij} > 0.784$

Group 3 (items 3 and 6) $0.784 >= S_{ij} > 0.692$

Group 4 (items 3, 6 and 5) $0.692 >= S_{ij} > 0.463$

Group 5 (items 1, 2, 7 and 4) $0.463 >= S_{ij} > 0.212$

Group 6 (all items) $S_{ij} <= 0.212$

Sneath (1957), (in Gordon, 1981: 36; Anderberg, 1973: 41) introduced a complete link algorithm that is very simple and is done in one step from the proximity matrix. Depending on S_1:

1 and 2 join at level 0.998

7 join (1 and 2) at level 0.804

3 join 6 at level 0.784

5 join (3 and 6) at level 0.692

4 join (1, 2, and 7) at level 0.463

(5,6 and 3) join (1,2,7 and 4) at level 0.212

-The next example shows the results of applying FN to the sample data records which similarities have been calculated using the Modified Euclidean distance coefficient:

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ .085 & 0 & & & & & \\ 2.98 & 2.96 & 0 & & & & \\ 5.74 & 5.73 & 5.86 & 0 & & & \\ 4.7 & 4.74 & 3.96 & 6.97 & 0 & & \\ 4.5 & 4.54 & 3.44 & 6.75 & 4.22 & 0 & \\ 3.04 & 2.98 & 3.79 & 5.1 & 5.47 & 5.6 & 0 \end{pmatrix} \end{array}$$

Applying the FN to the items in $\mathbf{D_1}$ will result in the following groups:

1 and 2 join at level .085

3 join (1 and 2) at level 2.98

7 join (1, 2, and 3) at level 3.97

5 join 6 at level 4.22

(5 and 6) join (1,2,3 and 7) at level 5.6

4 join (1, 2, 3, 7, 5, and 6) at level 6.97

-The next example shows the results of applying FN to the sample data records which similarities have been calculated using the Euclidean distance coefficient:

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ 1 & 0 & & & & & \\ 11.3 & 10.4 & 0 & & & & \\ 19.6 & 19 & 16.8 & 0 & & & \\ 23.7 & 24.2 & 28.7 & 32.4 & 0 & & \\ 8.3 & 8.5 & 12.6 & 22 & 21.2 & 0 & \\ 28.3 & 27.3 & 20.2 & 19 & 41.2 & 30.5 & 0 \end{pmatrix} \end{array}$$

1 and 2 join at level 1

6 join (1 and 2) at level 8.5

3 join (1, 2, and 6) at level 12.6

4 join 7 at level 19

(4 and 7) join (1, 2, 6 and 3) at level 22

5 join (1, 2, 6, 3, 4 and 7) at level 41.2

-The next example shows the results of applying FN to the sample data records which similarities have been calculated using the Canberra metric coefficient:

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ .008 & 0 & & & & & \\ 1.798 & 1.790 & 0 & & & & \\ 2.425 & 2.418 & 3.522 & 0 & & & \\ 3.509 & 3.517 & 1.854 & 2.331 & 0 & & \\ 2.672 & 2.680 & 1.018 & 4.485 & 1.514 & 0 & \\ 1.231 & 1.224 & 2.887 & 1.373 & 2.714 & 3.881 & 0 \end{pmatrix} \end{array}$$

1 and 2 join at level .008

3 join 6 at level 1.018

7 join (1 and 2) at level 1.231

5 join (3 and 6) at level 1.854

4 join (1,2 and 7) at level 2.425

(3,6 and 5) join (1,2,7 and 4) at level 4.485

-*Summary of FN*:

The results of applying the FN to the different proximity measures are depicted in the following Dendrogram figures (5-5: 5-8). Section (5-7-12) analyzes these results in detail.

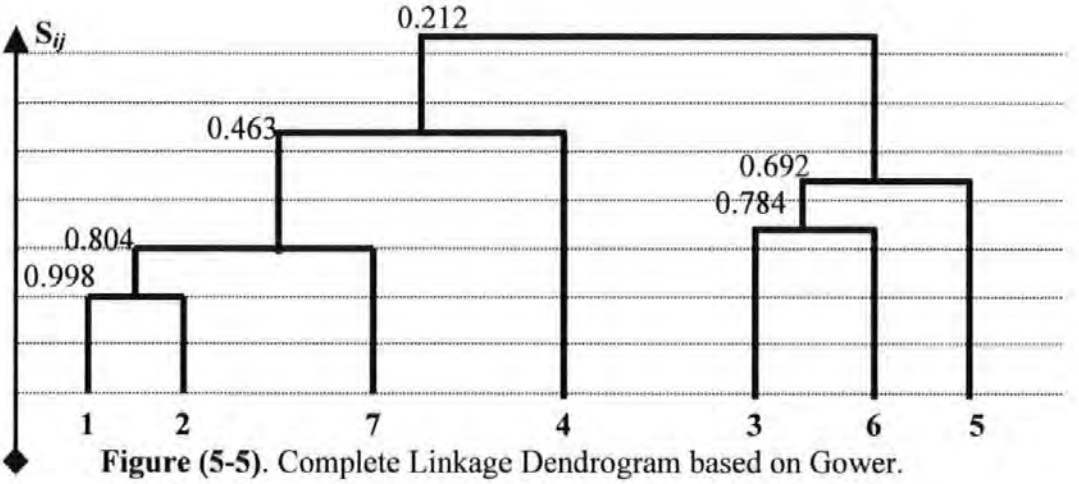


Figure (5-5). Complete Linkage Dendrogram based on Gower.

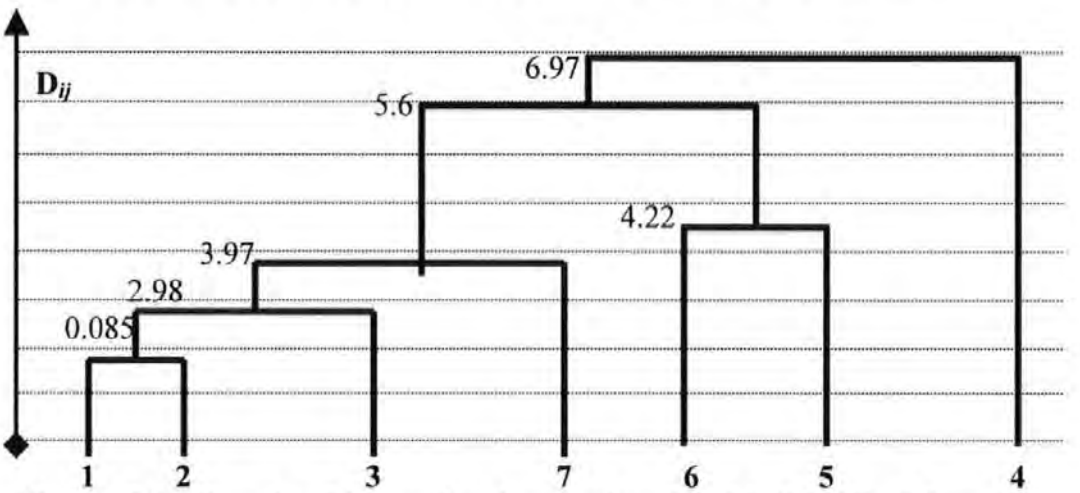


Figure (5-6). Complete Linkage Dendrogram based on Modified Euclidean.

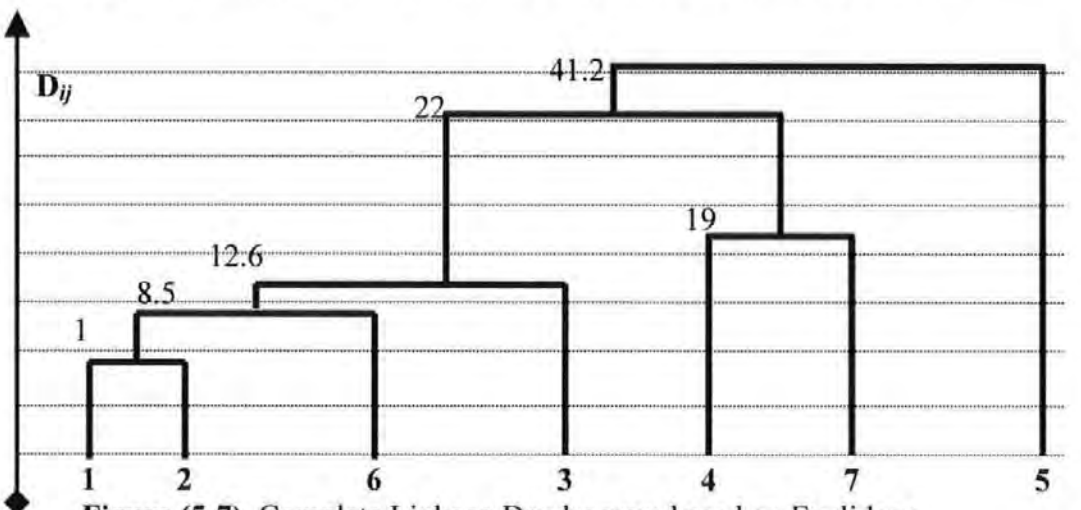


Figure (5-7). Complete Linkage Dendrogram based on Euclidean.

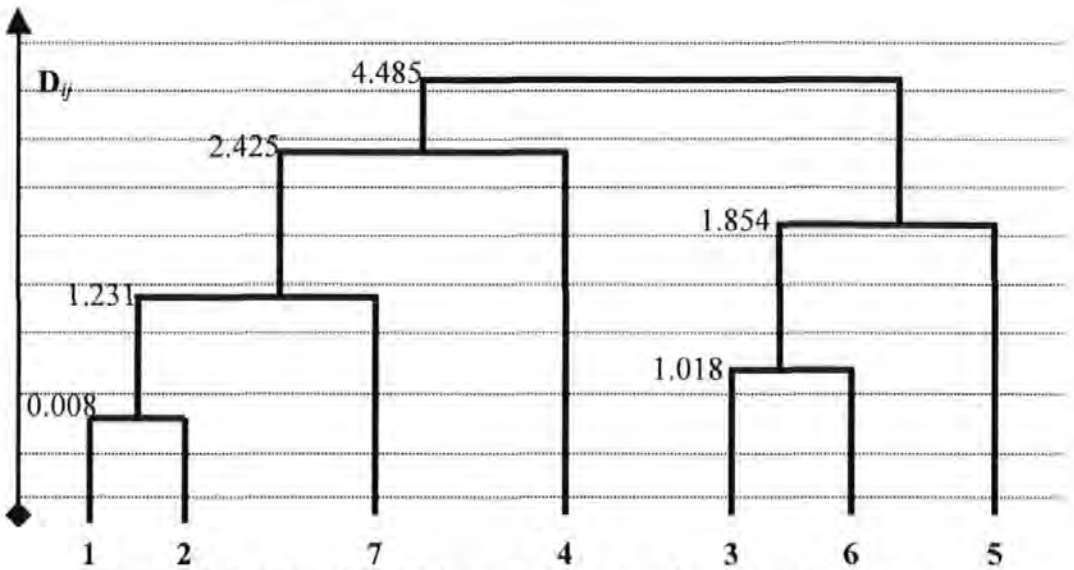


Figure (5-8). Complete Linkage Dendrogram based on Canberra.

3- Centroid Cluster analysis (in Gordon, 1981: 35; Everitt, 1980: 28; Anderberg, 1973: 140)

Groups are concentrated to lie in Euclidean space, and are then replaced by the co-ordinates of their centroids. The distance between groups is defined as the distance between their centroids.

A disadvantage of this technique is that when the two groups to be fused together have very different values in their variables the centroids of the new group will be very close to the larger group. Moreover, this cluster analysis technique focuses on the clustering of variables rather than clustering of items/records (Everitt, 1980), and requires a high computational power as the centroids are calculated after every fusion. Therefore, the Centroid Cluster analysis technique will not be used in the data used by this research.

4- Ward's method (in Wishart, 1998: 259; Gordon, 1981: 39; Everitt, 1980: 31; Anderberg, 1973: 142)

While some scholars have referred to this method as *Ward's* (Everitt, 1980; Anderberg, 1973), others have used the term *increase in sum of squares (ISS)* to refer to it (Wishart, 1998; Gordon, 1981). It assumes that group fusions results in a loss of information at each

stage of the analysis. This information loss can be measured by the total sum of squared deviations for every point from its mean of the cluster it belongs to. Each stage fusions are done for those who have the minimum increase in the error sum of squares (E.S.S). Ward can be calculated in two ways:

1. According to Gordon, 1981; Everitt, 1980; Anderberg, 1973 Ward's method is calculated based on the original data to cluster the data directly without the need for any proximity matrix. Cluster data is useful when the data set is extremely large e.g. over 50.000 records. Applications that use such large data sets include molecular biology and large survey analysis. The formula to calculate E.S.S is as follows:

$$E_p = \sum_{i \in p} \frac{\sum_j \left(X_{ij} - \mu_{pj}\right)^2}{V}$$ where j is a variable, X_{ij} is the value of j in record i, μ_{pj} is the mean of cluster p, and V is the number of variables. OR, the following formula is to be used when the variables are weighted:

$$E_p = \sum_{i \in p} \frac{\sum_j W_j \left(X_{ij} - \mu_{pj}\right)^2}{\sum_j W_j}$$ where j is a variable, X_{ij} is the value of j in record i, μ_{pj} is the mean of cluster p, and W_j is a differential weight (normally 1);

According to Everitt the error sum of squares (E.S.S.) is determined by the following formula (the formula assumes that the record is characterized by one attribute/variable):

$$\text{E.S.S} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

For example, if R_1 and R_2 are two records and X_i is a variable, then:

	X_i	X_i^2	
R_1	1	1	
R_2	2	4	
	3	5	E.S.S = 5 – ½*9 = 0.5

However, calculating Ward's based on the data to cluster depends on the variable values assigned for each individual rather than the proximity matrix. So, clusters found depend largely on the variable values not similarity or distance metric measures.

2. According to Wishart, 1998; Wishart, 1999b Ward's method can be used to cluster the data based on Euclidean distance proximity measure. Ward's based on a Euclidean metric measure proceeds as follows:
    a. Compute the Euclidean metric proximity matrix for all records in the sample;
    b. Start the fusions by combining the two records that will result in the minimum E.S.S, these two records are initially those whose distance D_{ij} is the minimum in the matrix;
    c. Compute the E.S.S;
    d. Combine the two clusters p and q that will result in the minimum of E.S.S. Where the total E.S.S for all clusters = $\sum_p E_p$, the two clusters to be combined should achieve the minimum of E.S.S = $E_{p\cup q} - E_p - E_q$;
    e. Repeat step d to combine cases whose union results in the minimization of E.S.S;
    f. Stop when all records are grouped in one cluster.

-The next example shows the results of applying Ward's to the sample data records actual variable values (i.e. without applying any proximity measure):

The results are based on the sample data records found in table (5-3).

Group fusions work in the following sequence:

1 join 2

6 join (1 and 2)

3 join (1, 2, and 6)

4 join 7

5 join (1, 2, 6, and 3)

(4 and 7) join (1,2,6,3 and 5)

-The next example shows the results of applying Ward's to the sample data records which proximities have been calculated using Euclidean metric measure:

$$\mathbf{D_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ 1 & 0 & & & & & \\ 11.3 & 10.4 & 0 & & & & \\ 19.6 & 19 & 16.8 & 0 & & & \\ 23.7 & 24.2 & 28.7 & 32.4 & 0 & & \\ 8.3 & 8.5 & 12.6 & 22 & 21.2 & 0 & \\ 28.3 & 27.3 & 20.2 & 19 & 41.2 & 30.5 & 0 \end{pmatrix} \end{array}$$

Group fusions work in the following sequence:

1 join 2

6 join (1 and 2)

3 join (1, 2, and 6)

4 join 7

5 join (1, 2, 6, and 3)

(4 and 7) join (1,2,6,3 and 5)

-<u>Summary of Ward</u>:

The results of applying Ward to the different proximity measures are depicted in the following Dendrogram figures (5-9:5-10). Section (5-7-12) analyzes these results in detail.

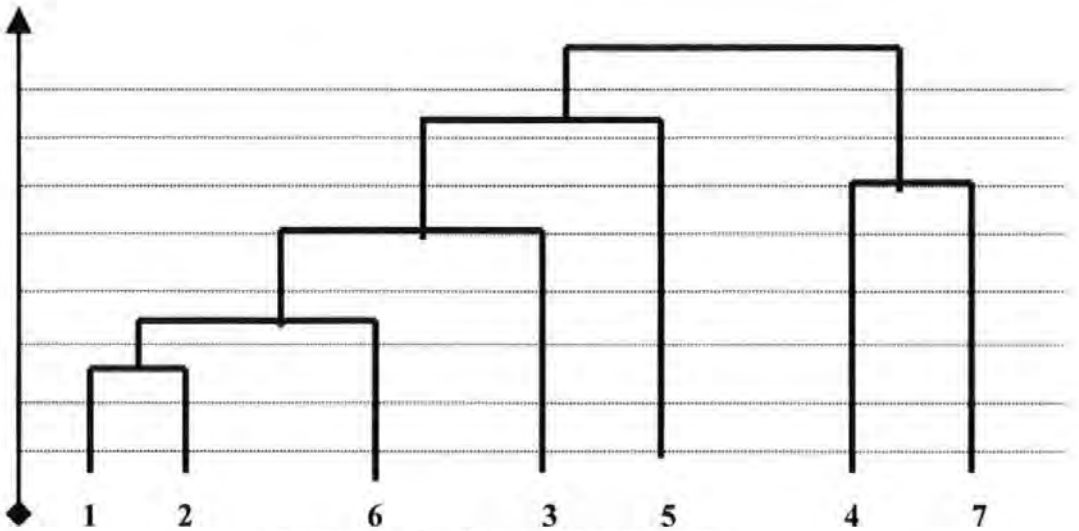


Figure (5-9). Ward's Dendrogram.

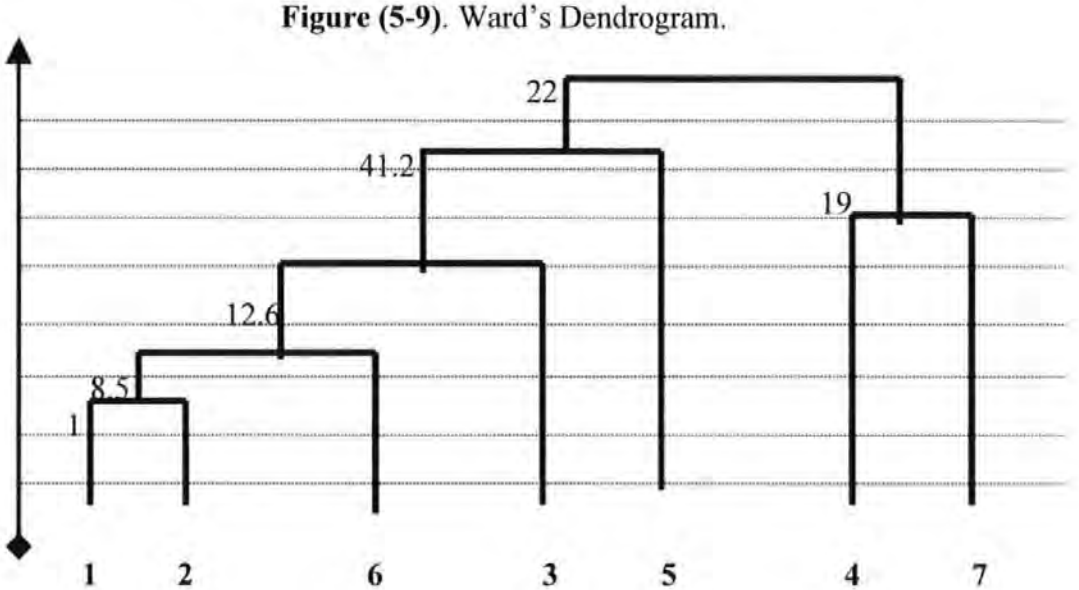


Figure (5-10). Ward's Dendrogram based on Euclidean.

Ward's method will not be applied to the other proximities (Modified Euclidean, Gower, and Canberra) because there is no evidence in literature of applying Ward's using any of these proximities.

5-7-11-2 Hierarchical Clustering techniques: II- Divisive techniques

The task undertaken by divisive methods is to split the initial set or items into two. For *n* individual items there are $2^{n-1} - 1$ ways to split them. These methods are generally used with very small data sets and even then require a large amount of computational power. Two types of divisive methods exist:

-*Monothetic* techniques which are based on the possession or not of a single specified attribute;

-*Polythetic* techniques which are based on the values taken by all attributes (Wishart, 1998).

-These techniques will not be used for the data used by this research because:

1. They are only of practical use for small data sets, whilst the data set used by this research is 2000 records;
2. These techniques are usually used on binary data sets. However, the data set used in this research is a mixture of data types;
3. Some of the *polythetic* techniques are designed to use a metric proximity measures. For example the most frequently used *polythetic* technique proposed by MacNaughton-Smith et al. (1964) depends on the Euclidean distance measure (Everitt, 1980);
4. Some of these techniques (i.e. The Automatic Interaction Detector Method-A.I.D.) are used for determining the variables and their categories that combine different groups with respect to some dependent variables.

5-7-11-3 Optimization Clustering techniques

This group of techniques differs from the hierarchical techniques in the sense that optimization techniques apply a relocation of the items/records which allows that a wrong-initial partition to be corrected at a later stage. Another difference between the optimization and hierarchical techniques is that the optimization techniques assume that the number of output groups (i.e. Clusters) has been defined a *priori* by the analyst. Some of the optimization techniques allow the number of groups to be altered later in the analysis. Problems associated with these techniques emerge when deciding on the number of clusters to be retained (Everitt, 1980).

There are a number of optimization techniques that are different with regard to: *methods of identifying the initial data partition*, and the *clustering criteria*.

Methods of identifying initial data partition. The majority of the techniques assume that there are k points in the p dimensional space, where k is the number of clusters. For instance, sometimes the first k points in the sample are taken as the initial k cluster mean vector, or the k points that are those furthest apart and these are used as initial cluster center. Then items are allocated to the cluster to whose center they are closest based on Euclidean metric measures (the following example will illustrate these ideas by applying the concepts to the 7 records sample).

The clustering criteria. (in Everitt, 1980: 42) Based on the clusters found, items reallocation between clusters is done to improve the clusters. This reallocation process continues until no further improvements can be found. Many clustering reallocation criteria (minimization of Trace[4] (W), minimization of the determinant of W, and maximization of Trace (BW^{-1})) are found; most of them are based on the following formula:

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

Where T is the dispersion matrix, W is the within-groups dispersion matrix, and B is the between-groups dispersion matrix. That is $W = \sum_{i=1}^{k} W_i$, W_i is the dispersion matrix for cluster i.

The following example applies the optimization technique (Minimization of Trace (W) which is the most frequently used in research according to Everitt (1980)) based on a Euclidean metric measure between the 7-student records sample.

$$= \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ 1 & 0 & & & & & \\ 11.3 & 10.4 & 0 & & & & \\ 19.6 & 19 & 16.8 & 0 & & & \\ 23.7 & 24.2 & 28.7 & 32.4 & 0 & & \\ 8.3 & 8.5 & 12.6 & 22 & 21.2 & 0 & \\ 28.3 & 27.3 & 20.2 & 19 & 41.2 & 30.5 & 0 \end{pmatrix} \end{array}$$

[4] Trace of a Matrix is the sum of the elements on the main diagonal.

According to the previous Euclidean metric matrix, furthest distance between the matrix items is 41.2 that is the distance between 5 and 7. So, the analysis starts at stage 1 with two groups:

	G 1	G 2
Stage (1):		
Items	5	7
M.V.[5]	(1,0,92,92,54,2,97,6,90,11,0)	(1,0,91,92,91,9,98,1,74,11,1)
Stage (2):		
Items	5	1, 7
M.V.	(1,0,92,92,54,2,97,6,90,11,0)	(1,5.5,91.5,92,78,8.5,98,1,74,11,1)
Stage (3):		
Items	5	1, 7, 2
M.V.	(1,0,92,92,54,2,97,6,90,11,0)	(1,7.3,91.7,92,74,8.3,98,1,74,11,1)
Stage (4):		
Items	5	1, 7, 2, 3
M.V.	(1,0,92,92,54,2,97,6,90,11,0)	(1,8.2,91.7,92,74.2,8.2,97.7,2.2,74.2,11,0.7)
Stage (5):		
Items	5	1, 7, 2, 3, 4
M.V.	(1,0,92,92,54,2,97,6,90,11,0)	(1,6.6,91.2,91.4,74.6,8.6,97.2,2.2,73.6,9,0.8)
Stage (6):		
Items	5	1, 7, 2, 3, 4, 6
M.V.	(1,0,92,92,54,2,97,6,90,11,0)	(1.2,7.3,91.3,91.5,72.8,7.5,97,2.8,73.8,9.3,0.7)

By now, the two cluster groups are G1 which contains item 5, and G2 which contains items 1, 2, 3, 4, 6, 7.

Trace (G 1) = Zero

Trace (G 2) = 881.67

[5] M.V.= Mean Vector

Now the trace of within-group matrix is:

Trace (G) = Trace (G 1) + Trace (G 2)

= 0 + 881.67 = 881.67

Now, consider moving item 3 to G 1, and recalculate the M.V. of the group and the trace, if the total trace (G) is decreased the item is kept in the new group and iterative process continues until no further improvements can be made.

Stage (7):

Items 5, 3 1, 7, 2, 4, 6

M.V. (1,5.5,92,92,64.5,5,97,6,82.5,11,0) (1.2,6.6,91.2,91.4,72.4,7.4,97,2.2,73.6,9,0.8)

By now, the two cluster groups are G1 which contains item 5, 3 and G2 which contains items 1, 2, 4, 6, 7.

Trace (G 1) = 411.5

Trace (G 2) = 841.2

Now the trace of within-group matrix is:

= 411.5 + 841.2 = 1252.7

Since the move of item 3 to G 1 has caused an increase in total trace, the item is not moved from its original cluster in G 2.

- *Important notice*: This optimization technique (Minimization of Trace [W]) has been applied to the same data set using the modified Euclidean metric measure, to see whether or not it will give different results if applied to the two proximity measures. The results were the same ending up with two groups; G1 contains item 5, whilst group 2 contains items 1,2,3,4,6, and 7.

5-7-11-4 Density Search Clustering techniques

This group of clustering techniques assume that the items/records are distributed in two clustering areas one is a high density area, and the other is of low density. Several techniques in this group are based on the NN and FN hierarchical techniques. This group

of techniques has emerged to overcome the main problem of hierarchical techniques that is *chaining* (Everitt, 1980). Chaining is the tendency of the technique to incorporate entities into an existing cluster rather than initiating new clusters. There are many techniques that are classified as density search. For example TAXMAP, CARTET Count, Detecting Fuzzy Sets, and Mode analysis. However, due the criticism and lack of usage of the last three techniques (i.e. mode analysis technique fails to detect both large and small clusters concurrently, Fuzzy Sets technique depends on Euclidean Distance measure), the TAXMAP technique will be applied to the sample data set.

1- TAXMAP clustering technique (Carmichael and Sneath, 1969 in Everitt, 1980: 47)

This technique was introduced by Carmichael and Sneath (1969). Clusters are found using the FN hierarchical technique. The difference between the two is that this technique uses criteria to judge when to stop adding new items/records to the cluster. That is not to add the new item if the expected item is much further than the last item admitted depending on the discontinuity in closeness. Carmichael and Sneath (1969) used a measure obtained by subtracting the drop in the average similarity on addition of an item to the cluster from the new average; the measure decreases until there is a big *drop* in the average similarity.

$$\mathbf{S_1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ .998 & 1 & & & & & \\ .759 & .761 & 1 & & & & \\ .463 & .466 & .368 & 1 & & & \\ .525 & .523 & .718 & .267 & 1 & & \\ .591 & .589 & .784 & .212 & .692 & 1 & \\ .804 & .806 & .611 & .586 & .511 & .395 & 1 \end{pmatrix} \end{array}$$

Based on the results of FN, the most two similar items in the matrix are items 1 and 2 where the similarity is 0.998. The next point is to add the closest point to the cluster that is item 7. The average similarity is now between the items of the clusters:

Cluster members	Prospective item
1, 2	7
Similarity 0.998	

Average similarity (1, 2, and 7) = 1/3 (0.998 + 0.804 + 0.806) = 0.869. Thus the drop in similarity would be 0.998 – 0.869 = 0.129. And the measure of discontinuity would be 0.869 - 0.129 = 0.74. This means that the cluster is now 1,2 and 7.

Low values (Less than 0.5) indicate that this item should not be added to the cluster. In cases whereas low values detected then the item should not be added and a new cluster is formulated.

Then, Consider adding items 3 and 6 to the cluster members, then:

Cluster members	Prospective item
1, 2, and 7	3
Similarity 0.869	

Average similarity (1, 2, 7 and 3) = 1/6 (0.998 + 0.804 + 0.806 + 0.759 + 0.761 + 0.611) = 0.789. Thus the drop in similarity would be 0.869 – 0.789 = 0.08. And the measure of discontinuity would be 0.789 - 0.08 = 0.709. This means that the cluster is now 1,2, 7 and 3.

Now, consider adding item 6:

Cluster members	Prospective item
1, 2, 7, and 3	6
Similarity 0.789	

Average similarity (1, 2, 7, 3 and 6) = 1/10 (0.591 + 0.589+ 0.784+ 0.395+ 0.998+ 0.759+ 0.804+ 0.761+ 0.806+ 0.611) = 0.709. Thus the drop in similarity would be 0.789 – 0.709 = 0.08. And the measure of discontinuity would be 0.709 - 0.08 = 0.629. This means that the cluster is now 1,2, 7, 3 and 6.

Now, consider adding item 5:

Cluster members	Prospective item
1, 2, 7, 3, and 6	5
Similarity 0.709	

Average similarity (1, 2, 7, 3, 6 and 5) = 1/15 (0.591 + 0.589+ 0.784+ 0.395+ 0.998+ 0.759+ 0.804+ 0.761+ 0.806+ 0.611+ 0.525+ 0.523+ 0.718+ 0.692 + 0.511) = 0.671. Thus the drop in similarity would be 0.709 – 0.671 = 0.038. And the measure of discontinuity would be 0.671 - 0.038 = 0.633. This means that the cluster is now 1,2, 7, 3 and 6.

Now, consider adding item 4:

Cluster members	Prospective item
1, 2, 7, 3, 6 and 5	4
Similarity 0.671	

Average similarity (1, 2, 7, 3, 6 and 5) = 1/21 (0.591 + 0.589+ 0.784+ 0.395+ 0.998+ 0.759+ 0.804+ 0.761+ 0.806+ 0.611+ 0.525+ 0.523+ 0.718+ 0.692 + 0.511+ 0.463+ 0.466+ 0.369 +0.267+ 0.212 +0.586) = 0.592. Thus the drop in similarity would be 0.671 – 0.592 = 0.079. And the measure of discontinuity would be 0.592 - 0.079 = 0.513. This means that the cluster is now 1,2, 7, 3, 6, 5, and 4.

5-7-11-5 Clumping Clustering techniques

The previous groups of classification techniques end up with a number of disjoint clusters. However, in some other application areas the clusters need to be overlapping. Clumping techniques have been used in the area of language studies and disease diagnosis. Whereby one word has more than one meaning and one patient might carry more than one disease at a time. These techniques will not be considered for application here where each student' record belong to only one cluster at a time[6], so this group of techniques is irrelevant here. For details of these techniques refer to (Everitt, 1980).

[6] The decision that each student belongs to one cluster at a time given that these clusters reflect the student academic performance is based on discussions with schools' officials.

5-7-12 Discussion on the results of the clustering techniques

1. The analyst must know that there are many clustering techniques available each of which has its own assumptions and gives different results if applied to the same data sets, the decision on choosing the clustering technique should be made in the light of the advantages and disadvantages of the chosen technique and the type of data to be classified (Aas, 1999; Wishart, 1998; Everitt, 1980);
2. Deciding on the number of clusters is a problem that is common to most of the clustering techniques. A large body of research has not succeeded to find a satisfactory solution for determining the optimum number of clusters, and it largely depends on the problem being resolved (Aas, 1999; Gordon, 1981; Everitt, 1980);
3. From the NN dendrograms (Figures 5-1:5-4) we can see that:
    a. The output result is affected by the proximity measure used;
    b. Gower and Modified Euclidean are in close agreement in their results;
    c. Euclidean and Canberra are also in close agreement;
    d. The results of NN can be described as consistent because the technique always started with 1 and 2 (i.e. similar records). Moreover, 4 and 6 (i.e. dissimilar records) did not come in a consequent order;
    e. Chaining has appeared when using Modified Euclidean and Euclidean, with less effect when using Gower or Canberra;
4. From the FN dendrograms (Figures 5-5:5-8) we can see that:
    a. The output result is affected by the proximity measure used;
    b. Gower and Canberra are similar in their results;
    c. The results of FN can be described as consistent because the technique always started with 1 and 2 (i.e. similar records). Moreover, 4 and 6 (i.e. dissimilar records) do not come in a consequent order;

d. The FN is less affected by chaining than NN. For example applying the FN based on a Modified Euclidean measure is less affected by chaining than applying NN based on the same measure (i.e. Modified Euclidean);

e. Canberra metric measure produced the same results with both NN and FN;

5. From the Ward's dendrograms (Figures 5-9:5-10) we can see that:

   a. There is no difference in the results obtained when applying the technique without using any proximity measure or when a proximity measure is used;

   b. Ward's method either based on a proximity measure or on the data is less affected by chaining than NN and FN;

   c. The results obtained by Ward's is very logical and consistent because:

      i. Students 1 and 2 (i.e. similar records) started the fusions;

      ii. Students 4 and 6 (i.e. dissimilar records) did not come in a consequent order;

      iii. Students 3 and 5 (i.e. fairly similar records) came in a consequent order and they joined the students 1, 2, and 6, whilst they did not come in a consequent order neither in NN nor FN;

      iv. Students 5 and 7 (i.e. fairly dissimilar records) did not come in a consequent order, whilst they came in a consequent order in both NN and FN;

6. The best results have obtained by calculating Ward's based on a Euclidean metric proximity measure. That is, 1 and 2 (i.e. similar) came first, 5 came after 3 (i.e. fairly similar), 6 and 4 are away (i.e. dissimilar), and also 5 and 7 are away (i.e. fairly dissimilar), with less chaining effect;

7. Although the concept of hierarchical techniques was developed in biology, this group of techniques is now used in many areas. One advantage is that the question on the optimum number of clusters does not arise since the researcher is interested in the complete hierarchy. The biggest disadvantage associated with these techniques is their

inability to reallocate items, which might be misclassified at early stages (Everitt, 1980). However, others (Jardine and Sibson, 1968) said that the NN has the greatest mathematical appeal amongst, and would generate suitable results for most application areas;

8. The NN technique has the problem of *chaining* (Wishart, 1998; Everitt, 1980). Chaining means that the NN method tries to accumulate the new records/case on existing cluster(s) rather than creating new clusters. As a result a few clusters retain the majority of records whilst the remaining clusters have fewer number of records;
9. The optimization techniques suffer from a number of problems. The techniques are transformation dependent; that is different results would be obtained from applying the same technique to the same data set. However, the advantages of using optimization techniques are the ability to reallocate misclassified item in further stages, and these techniques also do not assume that all clusters are hyper spherical- have the same shape. The most serious problem with the optimization techniques is the large amount of computation power they require, which in turn makes them irrelevant for the very large data sets;
10. The density solutions suffer from the problem of sub-optimal solutions; they might be more than one solution for the data sets (i.e. maximum likelihood). TAXMAP also suffer from the problem of containing various parameters that control the technique and arbitrary chosen by the investigator (Everitt, 1980);
11. The Clumping techniques rather than their unsuitability for the data sets, they suffer from the problem of optimization techniques that is the computation power they need;
12. Wishart (1998) suggested the calculation of Ward's based on a Euclidean metric measure.

5-7-13 The proximity measure and clustering technique to be adopted by this thesis

Based on the previous discussion on the clustering techniques and the discussion of the proximity measures (Refer to section 5-7-9), Ward's method (I.S.S) will be used based on the proximity of Euclidean metric measure.

5-8 Discussion of the research objective No. 2-3 "The use of the KDD techniques within the ARDSS"

Based on the characteristics of the chosen data mining techniques, the related work reviewed in this chapter in the last sections and in chapter four (Refer to section 4-7), the following table (5-12) summarizes when to use each of the techniques.

The Data Mining technique	Goal (G)/ Task (T)	Use in related work
SQL	-(G): Description -(T): Summarization	-For data summarization -Finding shallow knowledge -Provide general statistics -Helping users to take structured decisions -Produce reports -Answers to FAQ
Visualization	-(G): Description -(T): Summarization	-For data summarization -Finding shallow and multi-dimensionasl knowledge -Has the presentational advantage
Clustering analysis	-(G): Description Prediction -(T): Clustering	-For data description and prediction -Used when data variables are inter-correlated

Table (5-12). The use of the KDD techniques.

Having discussed the components of the proposed DSS methodology (i.e. DSS, DW, and KDD including the data mining techniques), we will now develop the new DSS methodology and the tools for its implementation.

5-9 The information engineering (IE) approach

The basic idea of the Information Engineering (IE) approach is to bring the organisation's plans into the process of IS development. That is each IS development is derived from certain business requirements; the requirements are based on the organisation's goals and objectives. These goals and objectives and the business requirements drive the information systems development plans (Whitten, et al., 1994; Davids, 1992). There are two basic

elements of IE that distinguish it from other systems development approaches (e.g. prototyping, and end-user development):

1. *The development of the IS is closely aligned with the organisation's goals and objectives*. This is to provide the reason for developing the IS through linking the organisation goals and objectives to a particular IS. Other goals of this link are to provide the future analysts and designers with a clear understanding of the organisation-information system relationship into which their projects should fit, and to get the top management support and commitment to the IS (Davids, 1992);
2. *It views and analyzes the business as separate areas these areas are called the business areas*. Then design a business system for each business area. Each business area is defined in terms of its information needs and requirements and its processes. The information needs are then transformed into data entities and then making the link between the data entities and business processes. Then design the data structure, procedures, screens and reports (Whitten, et al., 1994; Davids, 1992).

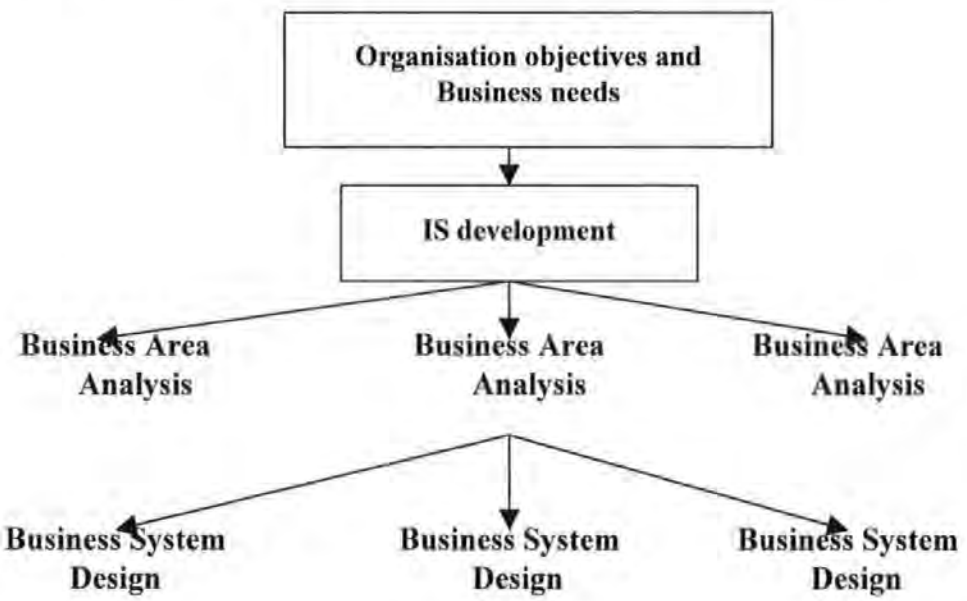


Figure (5-11). IE[7].

[7]Adapted from (Davids, 1992)

5-10 CASE tools

Computer Aided (or Assisted) Software Engineering (CASE) refers to the software used to assist in some or all phases of the SDLC (refer to chapter two-section 2-6-7-1). The basic feature of the CASE is the ability to generate an IS automatically based on the information stored about the business processes. CASE also offers a seamless transition from phase to phase with the same model throughout the SDLC (Davids, 1992). For example CASE is used to help in the business area analysis, user interface, business design, database design, and report generation. CASE tools also help automating some of the repetitive tasks during the SDLC for example drawing prototyping for screens and reports, generate program code, documentation, data flow diagrams (DFD), and checking design consistency (Laudon and Laudon, 2001; Hicks, 1993).

Integrated CASE (I-CASE) tools are special CASE types that can assist in all phases of the SDLC. Some of the CASE tools are supporting joint application development (JAD) in which a group of analysts, programmers and end-users can jointly design new applications (O'Brien, 1997).

5-11 The relationship between CASE tools and IE

Initially CASE tools were seen to be the implementation of the IE approach. CASE tools utilize the IE approach and present this approach to their users (Laudon and Laudon, 2001; Cool: Gen manuals, 1997; Davids, 1992). However, applying the IE approach is subject to the limitations of the CASE tools, since not all CASE tools adopt the full IE approach, whilst others completely adopt it e.g. I-CASE. Raghunathan (1996) mentioned that the use of CASE tool will significantly enhance the DSS development.

The CASE tool employed in this research is the Cool: Gen 5.0 by Computer Associates (CA). The objective of the Cool: Gen 5.0 (formerly *IEF* by TI, then *Composer* by

Sterling) software is to capture the information needs at the highest possible level of abstraction and transform them into executable application systems.

The researcher chooses to develop the model using CA Cool: Gen CASE tools because of the following characteristics (Baik, 2000; Cool: Gen manuals, 1997; Devlin, 1997):

- Cost savings achieved by using the Cool: Gen CASE tool;
- Developer productivity increases, by up to 300%;
- Dramatic improvements in business processes;
- Higher levels of customer satisfaction;
- Extraordinary flexible and high performance applications;
- Accelerated systems development;
- Applications ease of use, high greater growth potentials, and personalized solutions with built-in automated decision support;
- Supports most of the leading RDBMS (e.g. ORACLE, SYBASE, MS SQL, IBM DB Server), and many OS environments (e.g. UNIX, Windows NT, Windows 95/200, OS/390);
- The ability to generate code in different languages (e.g. C++, COBOL), which means that developers do not need to know a wide range of programming languages. They are only required to understand one simple English-like toolset, Cool: Gen;
- Users and developers of decision support systems implemented using the traditional approaches always having data availability, and data management problems.

5-12 The DW development

MS-SQL Server will be used to develop the data warehouse for the admission and registration function. MS-SQL Server supports both the relational and star schema structure models. Sørensen and Alnor (1999) said that MS-SQL Server can easily and efficiently handle both models (i.e. relational and star schema).

5-13 The proposed DSS methodology "Discussion of research objective No.2 "Develop a new DSS methodology"

Based upon the proposed DSS definition introduced in section (5-2), the justification of the use of CASE evaluated in section (5-10 and 5-11), the DW literature provided in chapter three, and the KDD literature provided in chapter four, the proposed DSS methodology is introduced now. The methodology is depicted in figure (5-12).

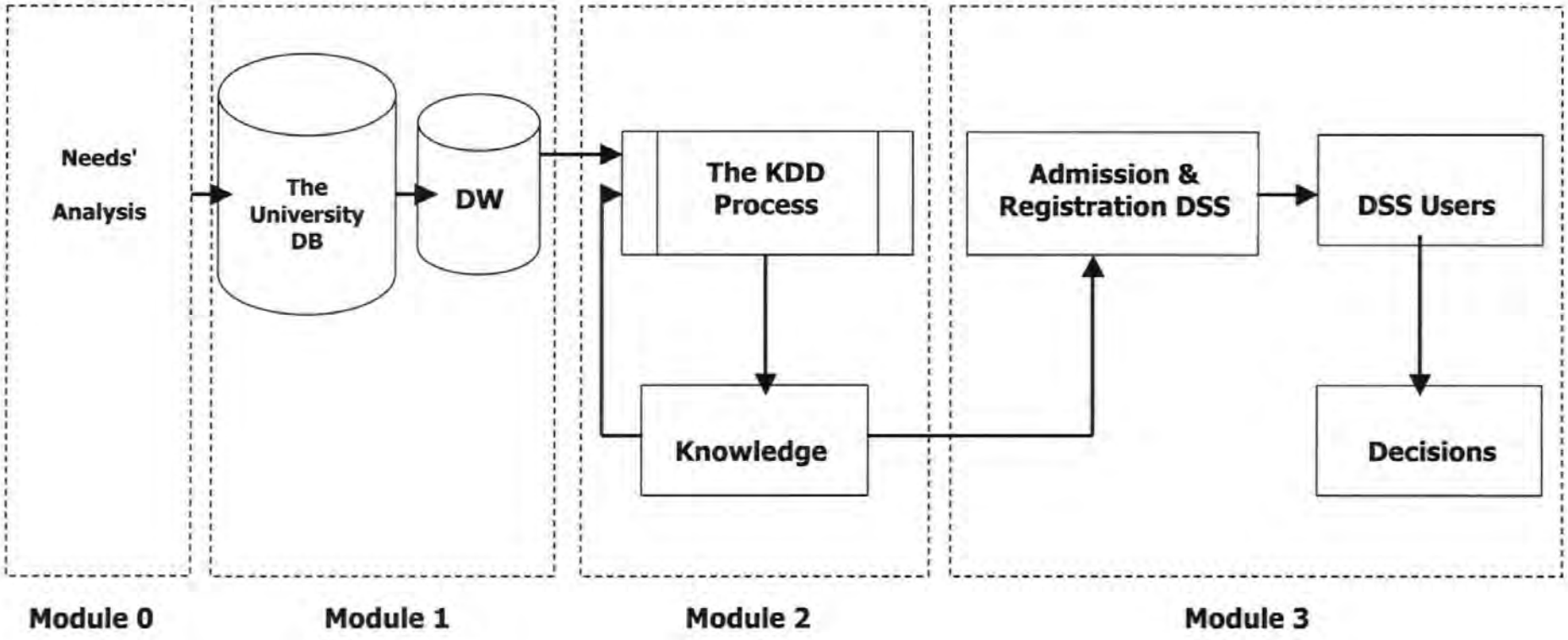


Figure (5-12). The proposed DSS methodology

The following table (5-13) summarizes the tools of each phase of the methodology.

The Module	Deliverable	Tools and Mechanisms
Module 0:	Needs' Analysis	-Questionnaire for user requirements -Cool: Gen CASE tools for analysis
Module 1:	Building the data warehouse	-MS-SQL Server: Star schema structure - Crystal Reports: Report generation tool.
Module 2:	Knowledge from the KDD process	-MS-SQL Server -Cool:Gen CASE tools -The data mining techniques are: SQL Visualization Clustering analysis
Module 3:	Building the DSS	-Cool: Gen CASE tools for development

Table (5-13). The methodology and its mechanisms.

Chapter summary

- A DSS is a computer-based information system that deals with semi-structured and unstructured problems facing managers at all management levels. The DSS goal is to enhance the decision quality and the manager effectiveness. To do so, the DSS integrates itself to the strategic data stores which is the data warehouse (DW), and to the knowledge discovery in database (KDD) process that will find the deep knowledge and hidden patterns in the DW and present them to the DSS user.
- SQL, Visualization, and Clustering analysis will be used as data mining techniques in this research.
- Ward's method based on a Euclidean proximity measure provided the best results when applied to the sample data records.
- CASE tool will enhance the development of the DSS.

- The proposed DSS methodology consists of four modules. Module zero handles the users' information needs. The first module is about DW, the second is about the KDD process and third is the integration of all DSS, DW, and the KDD together.
- Because of the importance of Cool: Gen CASE tool to the development of the DSS and to the thesis in general, the company profile is provided here. Computer Associates (CA):
    1. Computer Associates (CA) was founded in 1976 in the USA;
    2. CA achieved $ 6 Billion revenue in the fiscal year ended March 2000;
    3. CA products (i.e. Cool: Gen CASE tools, Business Intelligence, *Jasmine ii* eBusiness Solution; Enterprise Management) are used in 99 % of the Fortune 500 Companies;
    4. Companies in more than 100 countries in the world use CA products including:
        a. US Armed Forces (USA);
        b. Bank of Ireland "BOI" (Ireland);
        c. Royal Bank of Scotland "RBS" (UK);
        d. Enterprise Delivery Center "EDC" Munich (Germany);
        e. Scandinavian Garment Services "SGS" (Denmark);
        f. Japanese Systems Integrators Firm "JIEC" (JAPAN);
        g. Ministry of Defense (Egypt).

- The next chapter will study the current Admission and Registration IS in Egyptian Universities and will identify the users' requirements for a new Admission and Registration DSS (ARDSS) using a multi-part Admission and Registration DSS questionnaire (ARDSSQ). After that, chapter seven will start the development process of using the previous four-modules based methodology to implement the proposed ARDSS.

Chapter six

Extracting Users' Requirements

In order to build the proposed DSS two steps were undertaken. Firstly to evaluate the current Admission and Registration information systems in the Egyptian Universities. Secondly to extract the users' requirements for a new system. A multi-part questionnaire was developed and distributed to these Universities. This chapter starts by discussing the questionnaire development and then discusses the population and sample. After that it introduces the respondents distribution. Finally it analyzes the results and findings of the collected data according to the objectives of this dissertation.

6-1 Research objectives

The research objectives were stated in *chapter one* section 1-6. Some of these objectives are investigated using the Admission and Registration DSS research questionnaire (ARDSSQ). The remaining objectives will be investigated in *chapter seven.*

The research objectives which the ARDSSQ was developed to investigate are to:

1. Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:

    1-1 The managers' perspectives towards computers and their current Admission and Registration information systems

    1-2 Features of these information systems

    1-3 Functions of these information systems

2. Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:

    2-1 The managers' perspectives towards the role of computers and the ideal Admission and Registration information system

    2-2 The decisions that this DSS is expected to take

    2-3 DSS functions

    2-4 DSS characteristics

Each of the pre-stated objectives would be used as a questionnaire *construct*[1]. This means that the ARDSSQ will consist of *seven constructs*, each of which will investigate a single research objective.

Table (6-1) summarizes the relationship between the research objectives and the questionnaire constructs.

Objectives	Constructs
1-Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:	
1-1 The managers' perspectives towards computers and their current Admission and Registration information systems	1-1
1-2 Features of these information systems	1-2
1-3 Functions of these information systems	1-3
2-Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:	
2-1 The managers' perspectives towards the role of computers and the ideal Admission and Registration information system	2-1
2-2 The decisions that this DSS is expected to take	2-2
2-3 DSS functions	2-3
2-4 DSS characteristics	2-4

Table (6-1). Research objectives and questionnaire constructs.

6-2 Problem identification

This part of the research investigates the Admission and Registration functions taking place in the Egyptian Universities. The Egyptian Universities are classified by funding into two main categories; private and government funded.

[1] According to Cooper and Schindler (1998: 37) "A construct is an image or idea specifically invented for a given research and/or theory-building purpose".

All Egyptian Universities follow the same regulations and are controlled by the same authorities; Supreme Council of Universities and The Ministry of Higher Education. However, the Admission and Registration functions are different in these Universities in the sense of:

1. The Admission function.

Private Universities act independently when making decisions about accepting or rejecting students. However, in government Universities these decisions are taken centrally (there is a central board that accepts and distributes students between the different academic institutions in each University).

2. The Admission and Registration department structure.

In the private Universities, the Admission and Registration functions are centralized in one department for the entire University, whilst in government Universities there is a separate Admission and Registration department in each academic institute in the University.

Apart from the previous differences, the Admission and Registration functions in both University types are similar. They handle students' applications, map the students to the relevant academic institutions, doing course Registration and other grading-related jobs, graduation, class scheduling,...etc. Table (6-2) compares the two main categories.

Criteria	Government	Private
Number	Thirteen	Eight
Admission and Registration department structure	College/Faculty/Higher institute level	University level
Accepting/Rejecting Students	External decision	Internal decision
Students' file management	Internal	Internal
Registration functions	Internal	Internal

Table (6-2). Comparison between government and private Universities.

To investigate the decision makers' information needs relevant data needs to be collected. Due to the geographically dispersed locations of the Egyptian Universities, it was decided to use a questionnaire for this purpose.

6-3 The response base

According to the research objectives that the ARDSSQ is expected to investigate, the respondents are the decision makers in the area of Admission and Registration functions in Universities. They are:

1. Deans;
2. Deputy/Associate Dean for student academic affairs;
3. Registrars;
4. Admission officers;
5. Others who are entitled to do the same job under different job titles. E.g. Senior Academic Advisors.

6-4 The population and sample

6-4-1 Population

The population of this study is the Admission and Registration decision makers in the Egyptian Universities. These Universities are classified into two groups government and private-funded. According to the Egyptian Supreme Council of Universities statistics (1999), The UNESCO World List of Universities (2000), and The British Council Global Education and Training Information Service- Egyptian Universities (1999) there are twenty one Egyptian Universities; eight are private and thirteen are government.

6-4-1-1 The Private Universities

1. The American University in Cairo (AUC);
2. Arab Academy for Science and Technology and Maritime Transport (AASTMT);
3. 6th October University for Modern Sciences and Arts(MSA);

4. The Sixth October University;
5. MISR International University, Egypt (MIU);
6. MISR University for Technological Sciences (MUST);
7. Senghor University;
8. City University;

6-4-1-2The government Universities

9. Cairo University;
10. Alexandria University;
11. Ain Shams University;
12. Assiut University;
13. Tanta University;
14. Mansoura University;
15. Zagazig University;
16. Helwan University;
17. Minia University;
18. Menoufia University;
19. Suez Canal University;
20. South Valley University;
21. Al-Azhar University.

Each University consists of number of colleges, schools, faculties, and/or higher institutes and this is detailed in table (6-3) below:

The University	No of colleges/Faculties/Higher institutes
AUC	**3**
AASTMT	**8**
Senghor University	**4**
MSA	**8**
The Sixth October University	**7**
MIU	**6**
MUST	**6**
City University	**6**
Cairo University	**43**
Alexandria University	**27**
Ain Shams University	**17**
Assiut University	**16**
Tanta University	**21**
Mansoura University	**21**
Zagazig University	**28**
Helwan University	**18**
Minia University	**16**
Menoufia University	**18**
Suez Canal University	**22**
South Calley University	**16**
Al-Azhar University	**43**
Total	**354**

Table (6-3). Academic Institutions within the Egyptian Universities[2].

6-4-2 The sample

6-4-2-1The Sampling technique

As the number of Academic Institutions in Egyptian Universities is 354 distributed among *twenty-one* different Universities, a decision was made to target them all, and the Universities that will agree to participate would be used as a *sample*. All the *twenty-one* government and private Universities were contacted and asked to send their

[2] Source: Egyptian Supreme Council of Universities (1999), World List of Universities (2000), The British Council-Global Education and Training Infotmation Service (1999).

correspondence information and also to notify them that they will receive the questionnaire. Only *thirteen* Universities (*six private, seven government*) responded positively. Those Universities who did not respond were then contacted again but no reply received. After that the questionnaires were sent to those *thirteen* Universities.

6-4-2-2 The sample size

670 questionnaires were sent to the *thirteen* Universities that agreed to participate. Table (6-4) illustrates the number of questionnaires sent to each University.

The Universities	Questionnaires sent*
• **Private Universities:**	
1. The American University in Cairo (AUC)	**8**
2. Arab Academy for Science and Technology and Maritime Transport (AASTMT)	**18**
3. MISR International University, Egypt (MIU)	**10**
4. MISR University for Technological Sciences (MUST)	**14**
5. Senghor University	**14**
6. City University	**14**
Total private	**78**
• **Government Universities:**	
1. Alexandria University	**108**
2. Assiut University	**64**
3. Tanta University	**84**
4. Zagazig University	**112**
5. Menoufia University	**72**
6. Suez Canal University	**88**
7. South Valley University	**64**
Total government	**592**
Grand total	**670**

* For any private University the number of questionnaires sent = (number of academic institutes *2) + 2

* For any government University the number of questionnaires sent = (number of academic institutes) *4

Table (6-4). The sample size.

6-5 ARDSSQ – A measurement questionnaire for Admission and Registration IS in Egyptian Universities

The ARDSSQ is a multi-part questionnaire (*See Appendix "A"*). The first part of this questionnaire is designed to evaluate current Admission and Registration Information System in Universities. The second part of the questionnaire is designed to ascertain the information needs which are not satisfied by the current system and are required to build the ideal Admission and Registration DSS.

6-6 Questionnaire development

Churchill (1979) has introduced guidelines for the development of a new research instrument. Churchill emphasized the following structured approach:

1. *Specify the domain of the construct.* The domain of the questionnaire constructs should have a relationship with the research objectives that the questionnaire is expected to analyze. Preliminary meetings with respondents could be helpful at this stage;
2. *Generate sample items.* A sample questionnaire is to be drafted based upon the results of the preliminary meetings with respondents, telephone calls, reviewing documentation etc;
3. *Collect data.* After drafting the questionnaire it should be sent to a few respondents to comment on its items, wording, clarity of questions, structure and layout, terminology, and the required time to fulfill it;
4. *Purify the measures.* Based on the feedback received the questionnaire items, structure, wording, and terminology may be changed;
5. *Collect additional data.* When all the relevant changes requested by the reviewers have been made the final version is being mailed to all the respondents;
6. *Assess instrument validity and reliability.* At the end of the data collection phase the questionnaire validity and reliability have to be assessed;

7.*Develop norms.* The analysis of data, i.e. average and other statistics summarizing the distribution of scores.

The proposed approach by Churchill (1979) was found to be structured and solid. Chan, et al. (1998) followed this approach when developing their research questionnaire that was used to assess realized information systems strategy. However, *Churchill's approach could be improved by incorporating the following*:

- Churchill's approach is oriented for those who wish to introduce a new research instrument. However, related literature in the same field should be reviewed before his recommended steps. This suggested step would be of value to the researchers so that they could use other instruments as guidelines, and would be able to compare results;
- A pilot study is an important part of the research process for refining the research instrument therefore it should be included;
- A criteria should be established to determine a reasonable response rate;
- It is important to identify the research objectives, problem, and population.

Based on the previous discussion of Churchill's approach and the enhancements recommended above, the following research instrument development approach is proposed:

-A *pre-step* assumes that the research objectives, research problem, and the population have been clearly identified;

1. Stage one: Find a suitable questionnaire in the literature. In this step the literature is investigated for relevant issues that have been identified;
2. Stage two: Develop and validate the questionnaire-initial releases;
3. Stage three: Pilot study;
4. Stage four: The final release of the questionnaire;
5. Stage five: Questions' coding;

6. Stage six: Decide on the response rate;
7. Stage seven: Reliability and Validity of the questionnaire;
8. Stage eight: Data analysis and results.

The following table (6-5) summarizes the differences between Churchill's approach and the proposed approach.

Churchill's approach	The proposed approach
-	*pre-step* assumes that research objectives, research problem, and the population have been clearly identified
1- Specify the domain of the construct	1- Find a suitable questionnaire in the literature
2- Generate sample items	2- Develop and validate the questionnaire
3- Collect data	3- Pilot study
4- Purify the measures	4- The final release of the questionnaire
5- Collect additional data	5- Questions' coding
6- Assess instrument validity and reliability	6- Decide on the response rate
7- Develop norms	7- Reliability and Validity tests' scores of the questionnaire
-	8- Data analysis and results

Table (6-5). Comparison between Churchill's and the proposed approach.

The proposed instrument development approach is going to be discussed in detail and applied to the ARDSSQ in the next sections.

6-7 Stage one: Steps taken to find a suitable questionnaire in the literature

The development of a research instrument started by investigating the following sources to see if a relevant questionnaire existed.

1. Papers and articles were searched in the information systems periodicals including MIS Quarterly, Journal of Strategic Information Systems, Information & Management, and DSS.

2. E-mails were sent to some University professors in information systems in England, Ireland, USA, and Canada (University of Plymouth-England, Ulster-Ireland, Simon-Fraser-Canada, Georgia State University-USA, and others).
3. E-mails were sent to some Software Development companies in UK, Canada, and USA (SIS-Canada, Sterling[3]-UK, and others).
4. Conversations with Systems Analysis and Design professors in the School of Computing, Plymouth Business School.
5. Searching in the Instruments' Book (Stewart, et al., 1981); a two-volume book that includes all questionnaires used in Business Research. Only volume two was reviewed because volume one is all about measures of Satisfaction, whilst the reviewed volume two is about measure of Organisational Characteristics.
6. Searching the Internet to find a relevant questionnaire.

This research did not find a suitable questionnaire that dealt with DSS applied to Admission and Registration functions in Universities. Hence it was necessary to develop the required research instrument.

6-8 Stage two: Steps taken to develop and validate the questionnaire

1. Preliminary discussion sessions were held with both the Registrar and Admission Officer at the Arab Academy for Science and Technology and Maritime Transport (private University), and also with the former Registrar of the Faculty of Commerce, Alexandria University (government University).

At this stage the schools' officials were asked to raise any issues regarding their day-to-day work duties, main functions, the use of the Admission and Registration information systems, how satisfied with their systems' performance, and what are their information needs that require more attention.

[3] Sterling has recently been taken over by Computer Associates www.computerassociates.com.

Officials interviewed agreed that their work duties are classified into two main categories; *Admission-related* and *Registration-related.* Then they were asked to write down any possible problems, issues, and/or requirements that they would need an Admission and Registration information system to fulfill.

2. The first draft of the questionnaire was then prepared. Further sessions were conducted with schools' officials to comment onto the questionnaire, and changes (in wording, terminology, structure, questions addes/deleted) were made until an agreed format was reached.
3. Questionnaire review by other Ph.D. researchers. Five students at Nova South Eastern University in Florida, USA, reviewed the questionnaire. In this step they, they were asked to examine the content validity of the questionnaire, wording, sequence, and layout. The questionnaire items, direction, wording and layout were modified after this step.
4. Questionnaire review by IS Consultants. Two consultants were asked to comment on the questionnaire wording, sequence of questions, bias, and layout. One of them is responsible for the Admission and Registration information systems at the University of Plymouth. The second one is an IT Consultant (MCSE, MCT, MBA) in the USA. The questionnaire items, direction, wording and layout were modified after this step.
5. Questionnaire review by registrars and Admission officers in different schools. The goal of this step was to ensure that the questionnaire is understandable and unambiguous to practitioners. Some of the questions were re-phrased after this step.
6. Research Methodology professors reviewed the questionnaire. Based on their feedback the questionnaire scale has been changed in the last two questions (Q.24 and Q.25), originally these questions' followed a Likert scale (S.Agree, Agree, Neutral, Disagree, and S. Disagree), their advice was to change it to a dichotomous scale (Y/N or T/F) so that all questions would be of the same scale and are then statistically comparable.
7. The research supervisor reviewed the questionnaire after each of these steps.

8. Finally a reviewing step was carried out with some Registrars, Admission Officers, and an IT Consultant to ensure the validity of the final questionnaire. Up to this step seven revisions of the questionnaire were released before the final version was obtained.

6-9 Stage three: Pilot study

Before using the questionnaire to collect data it must to be piloted. In the literature the minimum number for a pilot varies starting from **5-10** according to many writers (Chan, et. al, 1998; Saunders, et. al, 1997; Fink, 1995b; Reynolds, et. al, 1993), to **50-100** according to others' (Green, et. al, 1988).

1. The questionnaire was disseminated to **30** senior Business Administration students, major MIS. They were asked to fill in the questionnaire and comment on it. They filled it in and no negative comments were received from them.
2. The questionnaire was disseminated to a group of **5** schools' officials to comment (some of whom were interviewed at the begining), no negative comments were received.

6-10 Stage four: The ARDSSQ

The final version of the ARDSSQ that was used for collecting data appears in Appendix (A). The ARDSSQ consists of 25 questions, investigating 7 objectvies (Refer to table 6-1). The relationship between each of the constructs and the questions is depicted in the following table (6-6).

Objectives / constructs	Abbreviation	Questionnaire questions
1-Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:		
1-1 The managers' perspectives towards computers and their current Admission and Registration information systems	**11 MPTCBIS**	**1 to 10**
1-2 Features of these information systems	**12 FEIS**	**11-1 to 11-6**
1-3 Functions of these information systems	**13 FUIS**	**12-1 to 12-8**
2-Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:		
2-1 The managers' perspectives towards the role of computers and the ideal Admission and Registration information system	**21 MPTICBIS**	**13 to 22**
2-2 The decisions that this DSS is expected to take	**22 DSSDE**	**23-1 to 23-23**
2-3 DSS functions	**23 DSSFU**	**24-1 to 24-11**
2-4 DSS characteristics	**24 DSSCH**	**25-1 to 25-11**

Table (6-6) . Objectives, constructs and the questions associated.

6-10-1 The ARDSSQ languages

The ARDSSQ was originally developed in English language. Since the ARDSSQ was developed to be sent to the Universities in Egypt, where in some of them Arabic is their first language, whilst in others English is their first langugae, a translation had to occur.

To ensure that there is no difference between the English and Arabic versions of the ARDSSQ, a *back translation technique* was carried out. The back translation technique was suggested by (El-Kot, 2001; Newmark, 1988; Hui and Triandis, 1985; Brislin, 1976; Brislin, 1970). The technique is used whenever a research instrument will be used in a multi-language population or sample.

The following steps were undertaken:

1. The ARDSSQ was developed and validated in English language;
2. The ARDSSQ was translated into Arabic language by the researcher, and checked by another Egyptian researcher;
3. The Arabic language version of the ARDSSQ was given to a translator in Egypt to translate it back into English language (*back translation*);
4. The two English copies were validated against each others by English-Arabic translator to ensure that there are no language differences (*back translation test).* Differences were identified and corrected;
5. The Arabic language was then validated by some Admission Officers and Registrars to ensure consistency and relevance of the terminology used.

The ARDSSQ Arabic versions were sent to those Univesities whose first language is Arabic, and the English versions to those Universities whose first language is English. Both copies are of equal number of pages (9 pages each).

6-11 Stage five: Questions' coding

For the general information questions both open-ended questions (e.g. name, address, telephone, e-mail) and category questions (e.g. type, position) were used. Where only the category questions will be part of the analysis. The main use of the open-ended question was for follow-up reasons.

The majority of the remaining questions are *list questions (Yes / No).* Where most of the questions asking about the availability or a certain element, feature, function, decision to take or not....etc. It has been established that *list questions* are the best type of question to use in these situations (Saunders, et. al, 1997; Teo and King, 1996). Open-ended questions

have been used to obtain personal opinion or for giving the respondents the right to add something which is originally not part of the list.

The data collected by this research was coded and **SPSS** 9.0 (The Statistical Package for the Social Sciences) was used to store and analyze the data.

For the General Information questions, the coding scheme is as follows:

Q.2: Government Universities coded 1, Private Universities coded 0.

Q.6: Dean 1, Associate Dean 2, Registrar 3, Admission Officer 4, Other 5.

For all other list questions (Yes/ No) the coding scheme is as follows:

Q.1 up to Q. 23: Yes coded 1, No coded 0. For those questions that respondents where asked to skip based on their answers to another question, code 9 was given. And finally code 99 for those questions without answers where no instruction was given to skip.

Questions (11, 12, 23, 24, 25) with extra space (open ended) where coded 1 which indicate Yes, and analysed separately in section (6-15) of this chapter.

6-12 Stage six: The Response rate

6-12-1 University-wise

The following table (6-7) illustrates the Universities that have already responded in relationship to the population and to the sample size as well.

The Universities	Part of the sample	Responded	Response Rate
• **Private Universities:**			
1-The American University in Cairo (AUC)	Y	Y	
2-Arab Academy for Science and Technology and Maritime Transport (AASTMT)	Y	Y	
3-6th October University for Modern Sciences and Arts(MSA)	N	-	
4-The Sixth October University	N	-	
5-MISR International University, Egypt (MIU)	Y	Y	
6-MISR University for Technological Sciences (MUST)	Y	Y	
7-Senghor University	Y	Y	
8-City University	Y	Y	
Total of 8 Private Universities	**Sample of 6**	**6**	**100 %**
Government Universities:			
9-Cairo University	N	-	
10-Alexandria University	Y	Y	
11-Ain Shams University	N	-	
12-Assiut University	Y	Y	
13-Tanta University	Y	Y	
14-Mansoura University	N	-	
15-Zagazig University	Y	Y	
16-Helwan University	N	-	
17-Minia University	N	-	
18-Menoufia University	Y	Y	
19-Suez Canal University	Y	Y	
20-South Valley University	Y	No	
21-Al-Azhar University	Y	No	
Total of 13 Government Universities	**Sample of 7**	**6**	**85.7 %**
Grand total	**13**	**12**	**92.3 %**

Table (6-7). The University-wise Response rate.

6-12-2 Respondent-wise

The following table (6-8) illustrates the respondents' distribution among the different academic institutions and the various positions followed by the response rate.

Institute	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level						5		7		35	**47**
Faculty of Science	4		5		3				3		**15**
Faculty of Commerce	2		2		5				1		**10**
Faculty of Law	3		1		4						**8**
Faculty of Hotels and Tourism	1		1		1						**3**
Faculty of Education	2		7		6						**15**
Faculty of Medicine	1		1		2						**4**
Faculty of Physical Education					2						**2**
Faculty of Dentist			1		1						**2**
Faculty of Pharmacy	2		2		1				1		**6**
Faculty of Veterinary	1				1						**2**
Faculty of Arts	1		2		2				3		**8**
Faculty of Agriculture	5		3		3				1		**12**
Faculty of Engineering	2		2		2	1		1		1	**9**
Faculty of Home Economics	1		1						1		**3**
College of Management				1		6		2			**9**
Faculty of Informatics	1										**1**
Faculty of Social Services	1		1		1						**3**
College of Marine Engineering						1					**1**
College of Maritime Studies						1					**1**
DSS Unit									6		**6**
Total respondents	**27**		**29**	**1**	**34**	**16**		**10**	**16**	**36**	**167**

Table (6-8). The distribution of respondents

Response rate = **167 / 670** = **24.9** %. This response rate is adequate in this kind of research that is based on mailed questionnaires (Chan, et. al, 1998; Saunders, et. al, 1997; Teo and King, 1996). Ravichandran and Rai (2000: 395) yielded 17.32 % response rate on their research on Quality Management in Systems Development, whilst Keil et al. (2000: 643) yielded 26 % on their survey on Escalated and Non-Escalated Software Projects.

The position contribution to the response rate is illustrated in table (6-9), whilst the University type contribution is illustrated in table (6-10).

Position	Responses	Per %
Dean	27	16.17
Associative Dean	30	17.96
Registrar	48	28.74
Admission Officer	10	5.99
Others; Director, Senior Academic Advisors	52	31.14
Total	**167**	**100 %**

Table (6-9). Response rate by position.

University	Responses	Per %
Government	**106**	**63.5**
Private	**61**	**36.5**
Total	**167**	**100 %**

Table (6-10). Response rate by University.

6-13 Stage seven: Reliability and Validity of the ARDSSQ

There are two major criteria in evaluating a measurement questionnaire: reliability and validity. In a recent study Boudreau, et al. (2001) have studied the "validation in information systems research", their research highlights the importance of validity and reliabiltiy of the instruments used in IS research. Table (6-11) summarizes the different thoughts and opinions on validity and reliability.

Scholars	Reliability	Validity	Measure
Bryan & Cramer (99)	-Test retest -Split half -Alpha	-Face -Concurrent -Construct -Predictive -Convergent -Discriminant	**Alpha, Construct-Factor or Bivariate analysis**
Rust & Golombok (99)	-Test retest -Parallel forms -Split half -Alpha	-Face -Concurrent -Construct -Predictive -Content	**Alpha, any other validity measure(s)**
Chan, et al. (96)	-Alpha	-Convergent -Discriminant	**Alpha, Convergent and Discriminant-Correlations**
Kline (93)	-Test retest -Split half -Alpha	-Face -Concurrent -Construct -Predictive	**Alpha, Construct-Factor analysis**
Nunnally (78)	-Test retest -Alternative forms -Split half -Alpha	-Construct -Predictve -Content & Face	**Alpha, Construct-Factor analysis**
Keil et al. (2000)	-Alpha -Conformity	-Convergent -Discriminant	**Alpha, Confomity Correlations**
Ravichandran & Rai (2000)	-Alpha -Werts Linn Josekog	-Convergent -Discriminant	**Alpha, P_c variance Correlations and X^2**

Table (6-11). Summary of the Reliability and Validity concepts.

In the following sections validity and reliability concepts will be introduced and discussed. The ARDSSQ reliability and validity tests scores will be illustrated in each section where relevant.

6-13-1 Reliability

Introduction. The reliability of a measure refers to its consistency. A measurement is reliable to the degree that it supplies consistent results. Reliability is a contributor to validity and is a necessary but not sufficient condition for validity (Bryman and Cramer, 1999; Cooper and Schindler, 1998; Saunders, et al., 1997; Kline, 1993; DeVellis, 1991; Nunnally, 1978). There are two types of reliability *internal* and *external*. External reliability refers to the degree of consistency of a measure over time. The process of

assessing external reliability by administering a test on two occasions (usually less than six months apart) to the same group of respondents, is called test-retest reliability. Internal reliability is concerned with whether each scale is measuring a single idea, and hence the items that make up the scale are internally consistent. Internal reliability is measured by split-half and Cronbach's alpha. Following evaluates these different reliability measures:

1-Test-retest.

This is used to measure the external reliability by administering the same test twice to the same subjects over an interval of less than six months to ensure the stability of results. Test-retest reliability is measured by correlation. Many problems are associated with the test-retest reliability process. For example interviewing events between the test and the retest may account for any discrepancy between the two set of results (Rust and Golombok, 1999; Bryman and Cramer, 1999). Further, participants may try to recollect their original answers which will create artificial consistency. The major defect is that experience in the first testing will usually affect the second testing (Nunnally, 1978). Nunnally (1978) does not recommend to use the retest as a measure of reliability unless there are cases where the respondent's answers will not be markedly affected by the first testing.

2-Parallel/Alternative-forms.

This involves administering the same test in an alternative or parallel form to the same person simultaneously or with a small delay. Consistency is again measured by correlation. Here we have two versions of the same test linked in a systematic way. Each person is given the two tests to complete and the reliability is obtained by calculating the Pearson Correlation coefficient. Parallel forms as a measure of reliability is rarely used (Rust and Golombok, 1999). Many reasons contributed to this, for example it requires twice the amount of work in preparing the two tests, and it carries many of the drawbacks of the test-retest (Rust and Golombok, 1999).

3-Split-half.

This establishes the degree to which instrument items are homogenous and reflect the same underlying construct(s). The items in a scale are divided into two groups either randomly or on an odd-even basis, and the relationship between respondents' scores for the two halves is computed. The agreement between the split-half is measured by the correlation coefficient. Split-half was criticized as being an over simple measure of reliability due to the criteria in associating each item with a group (Kline, 1993), and the correlation between the two halves will vary depending on how the items are divided (Nunnally, 1978). Rust and Golombok (1999) have stated that the reliability is always larger than the correlation between the two halves.

4-Cronbach's Alpha.

This calculates the average of all possible split-half reliability coefficients. Cronbach's Alpha is the most frequently used measure of internal consistency (Rust and Golombok, 1999; Chan, et al., 1998; Kline, 1993; Nunnally, 1978).

- Reliability of the ARDSSQ:

Major difficulties and criticism have been found in measuring a questionnaire's external reliability using test-retest or parallel forms and also with the over simplicity of the split-half as a measure of internal reliability. These are all factors that have contributed to the use of Cronbach's Alpha as the preferred and recommended measure of internal consistency by many scholars in information systems and other disciplines (Boudreau, et al., 2001; Chan, et al., 1998; Teo and King, 1996; Kline, 1993; Nunnally, 1978). A **0.7** is regarded by many researchers as a minimum figure for an adequate test score (Chan, et al., 1998; Kline, 1993; Nunnally, 1978), others said that **0.6** is a satisfactory score for exploratory research (Teo and King, 1996). In many situations, it is also recommended that the reliability test scores are to be calculated for the overall instrument and for each objective/construct as well (Chan, et al., 1998; Teo and King, 1996).

The ARDSSQ reliability Alpha is reported in table (6-12). The reliabilities associated with ARDSSQ are seen to be strong based on the previous minimum requirements.

The ARDSSQ overall Reliability Alpha coefficient is **0.96**

Objectives (constructs)	No of items	Alpha
11 MPTCBIS[4]	10	0.86
12 FEIS	6	0.95
13 FUIS	8	0.97
21 MPTICBIS	10	0.92
22 DSSDE	23	0.96
23 DSSFU	11	0.93
24 DSSCH	11	0.96

Table (6-12). Reliability Alpha coefficients.

6-13-2 Validity

Introduction. Validity refers to the extent to which a test measures the concept(s) that it intends or claims to measure (Bryman and Cramer, 1999; Rust and Golomok, 1999; Cooper and Schindler, 1998; Saunders, et al., 1997; Kline, 1993; DeVellis, 1991; Nunnally, 1978). *Unlike reliability, there is no single figure which indicates test validity* (Kline, 1993). Kline (1993) said that a test is valid for a particular purpose or with a particular group. Nunnally (1978) ascertained that one validates *NOT* a measurement instrument but rather a particular use of the instrument. It is evident that validity is a subjective issue and cannot be mesured by clear statistical methods (Bryman and Cramer, 1999; Kline, 1993; Nunnally, 1978).

There are many forms of validity, each of which has a various meaning and is measured differently:

1-Face validity.

This refers to the appearance of a test (Bryman and Cramer, 1999; Kline, 1993). Face validity concerns the acceptability of the test items, a test apparently reflects the contents

[4] Abbreviations of the constructs were introduced in table (6-6).

of the concept(s) in its questions (Rust and Golomok, 1999). Face validity is measured by judgmental methods (careful definition of the topic, items to be scaled, scale to be used). However, there is no necessary connection between face validity and true validity. The only demand that a test should be face valid is that without it respondents may not cooperate in the testing (Kline, 1993). Where face validity concerned with the instrument after it is being constructed, Nunnally (1978) recommended that face validity should be treated as part of the content validity which is concerned with the inspection of the final transformation from planning to constructing the instrument items.

2-Content validity.

This measures the extent to which the instrument provides adequate coverage of the topic under study (Rust and Golomok, 1999). Nunnally (1978) said that content validity requires a test to stand by itself as an adequate measure of what it supposed to measure. Content validity is measured by judgmental methods or panel discussion (using a panel of persons to decide how well the instrument meets the standards, correlating scores of different tests intending to measure the same thing). However, there are problems with ensuring content validity, obviously content validity rests entirely on appeals to reason regarding the adequacy with which the content has been cast in the form of test items (Nunnally, 1978).

- *Content validity of the ARDSSQ*:

1. The questionnaire content validity was measured by a panel discussion. Five PhD students at Nova South Eastern University in Florida, USA, reviewed the questionnaire. They were asked to examine the content validity of the questionnaire, wording, sequence, and layout. The questionnaire items, direction, wording and layout were modified after this step.
2. The same procedures were also used with the two IS consultants.

3-Predictive validity/Criterion-related validity.

A test is said to possess predictive validity if it can predict some relevant outcome (Rust and Golomok, 1999; Kline, 1993). Nunnally (1978) said that predictive validity is concerned when the purpose of an instrument is to estimate some variable that is external to it which is referred to as the criterion. It reflects the success of measures used for prediction or estimation, by identifying the degree to which the predictor is adequate in capturing the relevant aspects of the criterion. Predictive validity /Criterion-related validity is measured by correlation (comparing or correlating test scores to criterion scores- given that criterion scores are available). However, it is difficult to set up a good criterion to predict based upon which, this is why predictive validity is of little use (Bryman and Cramer, 1999; Kline, 1993; Nunnally, 1978).

4-Construct validity.

This attempts to identify the underlying construct(s) being measured and determines how well the test represents them (Bryman and Cramer, 1999). Nunnally (1978) said that to the extent a variable is abstract rather than concrete, we speak of it as being a *construct.* Construct validity is usually measured by factor analysis. *Construct validity is always the chosen measure of validity in many situations* (Rust and Golombok, 1999; Kline, 1993; Nunnally, 1978).

- *Construct validity of the ARDSSQ*:

1. The ARDSSQ was reviewed by registrars and Admission officers in different schools.
2. A final reviewing step was done with some registrars, Admission officers, and IT Consultant to ensure the validity of the questionnaire after the past modifications.

5-Concurrent validity.

To demonstrate concurrent validity, a test is correlated with another test of the same variable, both of which are administered concurrently (Rust and Golomok, 1999; Kline,

1993). Satisfactory concurrent validity requires a correlation of at least 0.7 between the two tests. However, concurrent validity is of little use because of many reasons. For example different test scales affect the correlations, many of the tests are being taken from another tests where it is very hard to compare between the two, and correlations would be difficult to interpret (Rust and Golomok, 1999; Bryman and Cramer, 1999; Kline, 1993).

6-Discriminant validity.

This implies a low levels of correspondence between a measure and other measures which are supposed to represent other concepts (Bryman and Cramer, 1999; Chan, et al., 1996; Campbell and Fiske, 1959). Discriminant validity is measured by bivariate analysis. It is also a recommended measure of validity by Chan e al. (1998), Keil et al. (2000), and Ravichandran and Rai (2000).

- *Discriminant Validity of the ARDSSQ*:

To assess the discriminant validity of the instrument objectives/constructs Chan, et al. (1998) recommended the inter-construct squared correlation coefficient matrix to be used, whilst Keil, et al. (2000) recommended the inter-construct correlation coefficient matrix. Keil, et al. (2000: 648) in their research followed the suggestion made by Ghiseli, et al. (1991) which stated that a correlation coefficients greater than **0.80** represent extreme cases. Wonnacott and Wonnacott (1990) also ascertained that high correlation coefficients are not recommended.

Chan, et al. (1998) measure of discriminant validity depends on factor analysis which is inappropriate to this research questionnaire where the scales used are all dichotomous (Yes/No). This means that we will adopt the approach used by Keil, et al. (2000), suggested by Ghiseli, et al. (1991), and fostered by Wonnacott and Wonnacott (1990).

Table (6-13) contains the inter-construct correlation coefficient, no cases were found to be extreme (> 0.80) in the matrix. The discriminant validity scores of the ARDSSQ are satisfactory.

The ARDSSQ Constructs	11 MPTCBIS	12 FEIS	13 FUIS	21 MPTICBIS	22 DSSDE	23 DSSFU	24 DSSCH
11 MPTCBIS	1						
12 FEIS	0.76	1					
13 FUIS	0.78	0.35	1				
21 MPTICBIS	0.13	0.22	0.23	1			
22 DSSDE	0.18	0.12	0.15	0.32	1		
23 DSSFU	0.10	0.10	0.15	0.38	0.43	1	
24 DSSCH	0.01	0.01	0.10	0.28	0.15	0.41	1

* P > .05 for the shaded cells, which represent insignificant correlations.

Table (6-13). Inter-construct Bivariate Correlations Matrix.

7-Convergent validity.

This attempts to demonstrate that each measure harmonizes with another measure (Bryman and Cramer, 1999; Chan, et al., 1996; Campbell and Fiske, 1959). This could be done by using different measures to see how far there is convergence, or by using observations in addition to the questionnaire (Jenkins, et al., 1975). It is a recommended measure of validity by Chan e al. (1998), Keil et al. (2000), and Ravichandran and Rai (2000).

- *Convergent Validity of the ARDSSQ* :

Table (6-14) shows the Item-to-Construct Correlations Matrix that is used to examine convergent validity. As indicated previously in discriminant validity this research follows the method proposed by Keil, et al. (2000) and Ghiseli, et al. (1991). The item measures should have highest correlation value with the construct it belongs to. When applying this rule to the results shown in table (6-13) we discover the following:

1. the first construct 11 MPTCBIS has significant correlations with the item measures (Q. 3: Q. 10) that are supposed to be part of it, whilst not significant with (Q. 1 and 2). The construct retains the highest correlations with the item measures (Q. 4, 5, 7: 10) that are designed to be part of it, whilst not the highest with (Q. 1:3, 6). Further

analysis shows that the insignificance of Q.1 and 2 are due to the Government Universities. When the correlations are calculated for each type of University it gives a closer value to the ones in the table for the Government, and 1 with the Private because all private Universities have computers and managers having computers on their disks, further they also running a CBIS. This also affects the significance of the correlation. Concerning Q. 3 the reason again is the University type. Finally, Q. 6 will be retained in the first construct because it still has a strong and significant correlation (.81), however, its highest correlation (.86) with the second construct;

2. For the remaining constructs (12 FEIS, 13 FUIS, 21 MPTICBIS, 22 DSSDE, 23 DSSFU, and 24 DSSCH), they all have significant correlations with all their item measures that are part of them. Also, all of them retain the highest correlations with all the item measures that are designed to be part of each of which;
3. Some item measures have weak correlations (Q. 13, 14, 21, 22, 23_2, 24_3) however, they are retained in their construct because of the following:

   -the high reliability scores (i.e. Alpha coefficient) of the constructs they belong to;

   -they are all having significant correlations;

   -these correlations are the highest.

Constructs	Item Measures	Constructs						
		11 MPTCBIS	12 FEIS	13 FUIS	21 MPTICBIS	22 DSSDE	23 DSSFU	24 DSSCH
11 MPTCBIS	Q. 1	**.07**	.04	.05	.06	.08	.10	.04
	Q. 2	**.05**	-.19	-.20	-.08	.11	.01	-.16
	Q. 3	**-.85**	-.90	-.85	-.22	-.13	-.08	-.03
	Q. 4	**.84**	.84	.84	.16	.15	.14	.09
	Q. 5	**.85**	.82	.84	.21	.17	.15	.07
	Q. 6	**.81**	.86	.84	.19	.13	.12	.05
	Q. 7	**.75**	.61	.66	.10	.08	.05	.01
	Q. 8	**.91**	.82	.82	.12	.14	.07	.04
	Q. 9	**.63**	.35	.35	-.07	.14	-.04	.05
	Q. 10	**.47**	.25	.23	-.11	.12	-.03	-.02
12 FEIS	Q. 11_1	.82	**.93**	.85	.19	.09	.02	.01
	Q. 11_2	.79	**.94**	.87	.22	.13	.11	.04
	Q. 11_3	.76	**.96**	.82	.24	.13	.06	.01
	Q. 11_4	.78	**.96**	.85	.21	.11	.09	.01
	Q. 11_5	.81	**.93**	.85	.23	.13	.04	.01
13 FUIS	Q. 12_1	.79	.85	**.94**	.21	.13	.09	.07
	Q. 12_2	.79	.87	**.93**	.24	.19	.16	.09
	Q. 12_3	.81	.84	**.92**	.25	.11	.09	.03
	Q. 12_4	.76	.82	**.92**	.23	.13	.11	.03
	Q. 12_5	.78	.83	**.94**	.24	.18	.16	.06
	Q. 12_6	.78	.78	**.94**	.21	.10	.14	.04
	Q. 12_7	.78	.81	**.93**	.21	.16	.13	.07
21 MPTICBIS	Q. 13	.13	.11	.20	**.31**	.14	.03	.16
	Q. 14	-.04	-.05	-.07	**.32**	.19	.19	.09
	Q. 15	.13	.13	.19	**.45**	.11	.19	.18
	Q. 16	-.14	-.08	-.04	**.38**	.10	.08	.08
	Q. 17	.05	.02	,06	**.40**	.09	.10	.10
	Q. 18	.15	.24	.21	**.64**	.31	.31	.26
	Q. 19	.12	.17	.19	**.52**	.23	.31	.29
	Q. 20	.08	.16	.21	**.42**	.19	.28	.24
	Q. 21	.02	-.02	.05	**.23**	.16	.19	.15
	Q. 22	-.02	-.05	.04	**.22**	.16	.21	.01
22 DSSDE	Q. 23_1	-.01	.02	.10	.12	**.50**	.17	.07
	Q. 23_2	.02	.04	.08	.23	**.37**	.21	.12
	Q. 23_3	.16	.17	.17	.27	**.50**	.17	.30
	Q. 23_4	.22	.17	.15	.23	**.44**	.19	.16
	Q. 23_5	.17	.19	.19	.33	**.48**	.34	.16
	Q. 23_6	-.01	-.05	.04	.15	**.42**	.13	.18
	Q. 23_7	-.01	-.05	-.04	.23	**.54**	.24	.24
	Q. 23_8	.05	.04	.06	-.02	**.50**	.15	.09
	Q. 23_9	.06	.07	.07	.18	**.56**	.26	.11
	Q. 23_10	.15	.13	.17	.36	**.61**	.35	.18
	Q. 23_11	.18	.18	.14	.35	**.55**	.33	.14
	Q. 23_12	.06	.01	.07	.18	**.52**	.23	.15
	Q. 23_13	.12	.13	.13	.20	**.51**	.26	.09

	Q. 23_14	.09	-.02	.03	.06	**.60**	.26	-.05
	Q. 23_15	.19	.07	.09	.11	**.62**	.27	-.06
	Q. 23_16	.16	.21	.25	.30	**.51**	.31	.13
	Q. 23_17	.20	.20	.19	.18	**.56**	.25	.07
	Q. 23_18	.03	.02	-.02	.17	**.55**	.25	.13
23 DSSFU	**Q. 24_1**	.09	.03	.17	.18	.42	**.66**	.32
	Q. 24_2	.03	.03	.08	.16	.34	**.57**	.24
	Q. 24_3	-.01	-.05	.04	.15	.29	**.32**	.19
	Q. 24_4	.07	.07	.15	.11	.25	**.40**	.37
	Q. 24_5	.11	.05	.22	.17	.27	**.47**	.28
	Q. 24_6	-.04	-.03	.03	.23	.32	**.55**	.30
	Q. 24_7	-.02	-.08	-.07	.24	.34	**.54**	.27
	Q. 24_8	.03	-.01	.01	.24	.37	**.37**	.17
	Q. 24_9	-.09	.05	.03	.20	.26	**.59**	.34
24 DSSCH	**Q. 25_1**	.17	.14	.27	.24	.27	.35	**.60**
	Q. 25_2	-.11	-.08	-.01	.13	.17	.28	**.56**
	Q. 25_3	-.07	-.08	-.04	.28	.20	.40	**.60**
	Q. 25_4	.11	.07	.20	.23	.29	.36	**.53**
	Q. 25_5	.10	.12	.25	.22	.32	.35	**.54**
	Q. 25_6	.07	.05	.13	.16	.24	.37	**.56**
	Q. 25_7	-.01	.01	.07	.16	.21	.33	**.53**
	Q. 25_8	.02	.03	.06	.16	.20	.39	**.62**
	Q. 25_9	.04	.02	.07	.14	.15	.27	**.55**

* P > .05 for the shaded cells, which represent insignificant correlations.

Table (6-14). Item-to-Construct Correlations Matrix.

6-14 Stage eight: Questionnaire Analysis

In the following sections, the data collected by the ARDSSQ will be analyzed. Each of the research objectives will be analyzed using an equivalent questionnaire construct (Refer to section 6-1 for details). *Each of the questionnaire constructs will be analyzed in terms of the following three dimensions*:

1. University type;
2. Respondent position;
3. Whether the University uses a CBIS or not.

The reason for using these three dimensions is due to the nature of the population (Refer to item 6-4), the response base (Refer to item 6-3), and the expected effects of using CBIS on the answers. And also to identify areas of commonality and discrepancy between the major segmentations identified within the population.

6-14-1 Discussion of the first objective

1-Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:

1-1 The managers' perspectives towards computers and their current Admission and Registration information systems	Questions **1: 10**

Table (6-15) shows the results of questions 1-10[5].

Discussion

- These questions investigate the managers' perspectives towards computers and their current Admission and Registration information systems.
- *Overall*: the majority of the respondents reported positively on nine questions (1: 8, and 10), whilst the negative and positive responses are equal with regard to question 9.
- *The University type dimension*: This dimension has an effect on the answers of two questions 3 and 9 (*Use of CBIS, Use of experience against abnormal system results*), whilst it has no effect on the remaining questions (1: 2, 4: 8, and 10), on which whether the respondent belongs to a Private or Governement University, it does not make any difference in his responses. The questions where the University type has effect on are 3 and 9 whilst the majority of Private Universities reported positively on both questions, the majority of Governement reported negatively on them. On the remaining questions of this construct (i.e. 1:2, 4: 8, and 10) the responses ranging from 57 to 100 %.
- *The respondent position dimension:* This dimension also has and effect on the answers of three questions 3, 9, and 10 (*Use of CBIS, Use of experience against abnormal system results, Mix the result and experience*), and has no effect on the answers of the remaining seven questions (1: 2 and 4: 8), on which either the repondent is a Dean, Associate Dean, Registrar, Admission Officer, or Other, this does not affect his responses to these questions. Responses to these questions range from 50 to 100 %. On

[5] Respondents were instructed to leave some questions without answer based on other their answers to a branching question e.g. if answer to question 3 was No, they were instructed to go directly to question 13.

question 3 (*Use of CBIS*) the majority reported positively, however Deans, Associate Deans, and Registrars reported negatively. On question 10 (*Mix the result and experience*) also the majority reported positively, however only Registrars reported negatively. On question 9 (*Use of experience against abnormal system results*) the responsednts are in-different to this question, however Associate Deans and Admission Officers reported negatively.

- *The use of CBIS dimension:* This dimension also has no effect on the answers to these questions. Whereas question 3 is a branching questions, that is starting from question 4 up to 10, all respondents who answered these questions should be using a CBIS (Because Non-users would have moved to question 13 after the branching question 3). Responses range from 61 to 99 %.
- On question 1 *(the admission and registration manager should have a computer on his desk),* the overall majority of respondents think that they should have a PC to perform their Admission and Registration work duties, neither of the three dimensions has effect on the responses to this question.
- On questions 2 and 3 (*do you have a PC, Use of CBIS*), the overall majority of responses reported positively on both. However, whilst neither of the dimensions has effect on the responses to question 2, both University type and Position affect the responses to question 3. The majority (92%) of Private Universities are using CBIS, and the majority (65%) of Government Universities are not using CBIS. On the other hand, the majority of Deans (56%), Associate Deans (53%), and Registrars (58%) are not using CBIS. Moreover, 100% of the Admission Officers are using CBIS.
- On question 6 (*Do you depend on the current information system to take decisions*), the overall majority (77% of the Private, and 75% of the Government Universities) reported positively, and neither of the dimensios has effect. This finding means that the current computerized systems are used to take decisions (i.e. DSS).

- On question 7 (*Do you encounter situations where your decision will be enhanced if you search in the students' history before making the decision*), the overall majority reported positively, and neither of the dimensions has effect. This means that adding the data warehouse component to an information system is expected to enhance the decision quality.
- Based on the previous discussion:
  - The managers' perspectives towards computers and their current Admission and Registration information systems is affected by these two dimensions; the University type and the managers' position.
  - The use of CBIS does not affect the managers' perspectives towards computers and their current Admission and Registration information systems.
  - The percentage of Private Universities that use CBIS (92%) is greater than the percentage of Government Universities that use CBIS (35%). However, the percentage of Admission and Registration managers who have PC's in Government Universities (72%) is greater than in Private Universities (67%). This means that PC's on the Government Universities managers's desks are used for other purposes e.g. word processors, e-mail, web access, etc.
  - Admission and Registration information systems are used extensively by Admission Officers more than the other types of managers (i.e. Deans, Associate Deans, Registrars, and Others). This could be due to any of the these reasons: the lack of enough time that these managers' need to spend on using CBIS, insufficient computer knowledge, they were not involved during the system's development, or the system does not meet their information and knowledge requirements.
  - The current computerized systems are used as DSS.
  - The DW component would enhance the decision quality.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.1 Having computer?	**No**	**5**		5		3	6		2	5	1
	Yes	**161**	100	95	100	97	94	100	98	95	99
Q.2 Have PC	**No**	**49**	33	28	15	10	35	10	46	39	22
	Yes	**117**	67	72	85	90	65	90	54	61	78
Q.3 Use CBIS?[1]	**No**	**74**	8	65	56	53	58		29		
	Yes	**93**	92	35	44	47	42	100	71		
Q.4 Use CBIS to All	**No**	**32**	32	41	50	23	30	20	43		36
	Yes	**58**	68	59	50	77	70	80	57		64
Q.5 IS linked to Historical Data	**No**	**17**	13	29	20	38	15	22	14		19
	Yes	**72**	87	71	80	62	85	78	86		81
Q.6 IS take decisions?	**No**	**22**	23	25		31	30		32		24
	Yes	**70**	77	75	100	69	70	100	68		76
Q.7 Decision Enhanced by History?	**No**	**11**	14	19	22		21		22		16
	Yes	**58**	86	81	78	100	79	100	78		84
Q.8 Use IS result and Update experience	**No**	**27**	25	36	42	46	32	20	22		30
	Yes	**64**	75	64	58	54	68	80	78		70
Q.9 Use experience	**No**	**13**	43	58	20	80	50	100	38		50
	Yes	**13**	57	42	80	20	50		63		50
Q.10 Mix result and experience	**No**	**5**	33	43		50	67	50			38
	Yes	**8**	67	57	100	50	33	50	100		62

Table (6-15). Objective 1-1 (The managers' perspectives towards computers and their current admission and registration information systems) Results.

[1] Q.3 is a branching question; respondents were instructed to go to Q.13 if the answer to Q.3 was No (I.e. they do not have CBIS). This means that those who have completed Q.4 to Q.12 use CBIS.

Discussion of the first objective (Cont'd)

1-Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:

1-2 Features of these information systems	Questions **11-1: 11-6**

Table (6-16) shows the results of questions 11-1 : 11-5[6].

Discussion

- These questions investigate the features of the current Admission and Registration information systems.
- *Overall*: the majority of the respondents reported positively on these two questions 11-1 and 11-5, whilst the majority reported negatively on the three reaining questions 11-2: 11-4.
- *The University type dimension*: This dimension has no effect on the answers to these questions. So, whether the respondent belongs to a Private or Governement University, it does not make any difference in his response; the majority of respondents in both University types reported either positively (i.e. questions 11-1 and 11-5) or negatively (i.e. questions 11-2: 11-4) on the same questions of this construct. The responses are ranging from 78 to 100% for the positive questions, whilst from 68 to 91% for the negative questions.
- *The respondent position dimension:* This dimension also has no effect on the answers to these questions. Either the repondent is a Dean, Associate Dean, Registrar, Admission Officer, or Other, he reported positively or negatively on the same questions of this construct. The responses are ranging from 79 to 100% for the positive questions, whilst from 50 to 95% for the negative questions.
- *The use of CBIS dimension:* This dimension also has no effect on the answers to these questions. All respondents who answered these questions should be using CBIS

[6] Questions with extra space (*Open-Ended*) like 11-6, 12-8, 23-19: 23-23, 24-10: 24-11, and 25-10: 25-11 will be discussed and analyzed in section 6-15 separately.

(Because Non-CBIS users would have moved to question 13 after the branching question 3). On the positive questions the responses were always 91%, whilst on the negative questions responses are ranging from 71 to 86%.

- On question 11-1 (*Printing reports that describe students' records*), 22% of the respondents in Government Universities reported negatively. Hence these systems do not have the ability to print description reports which is questionable because this is a basic IS feature. This implies that these systems are running as electronic data stores only (i.e. TPS), this thought is strengthen by the answer to question 11-5 (*It is an electronic store of students' data*) on which 92% of the respondents from Government Universities reported positively.
- Based on the previous discussion:
  - The features of the current Admission and Registration IS in the Egyptian Universities are not affected by any of these three dimensions; the University type, the respondent position, and the use of CBIS.
  - The Admission and Registration information systems in the Egyptian Universities have the following features:
    - Printing reports that describe students' records feature;
    - Electronic stores of students' data.
  - The Admission and Registration information systems in the Egyptian Universities do not have the following features:
    - Predicting the new applicants' performance;
    - Predicting the current-students' performance;.
    - Both description and prediction functions are available.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.11-1 IS feature Reports	**No**	**8**		22	8	21	10		6		9
	Yes	**83**	100	78	92	79	90	100	94		91
Q.11-2 IS feature Prediction	**No**	**79**	91	78	75	79	95	70	92		86
	Yes	**13**	9	22	25	21	5	30	8		14
Q.11-3 IS feature Prediction current	**No**	**65**	69	73	67	79	75	80	64		71
	Yes	**27**	31	27	33	21	25	20	36		29
Q.11-4 IS feature Description and Prediction	**No**	**65**	75	68	50	79	95	78	64		72
	Yes	**25**	25	32	50	21	5	22	36		28
Q.11-5 IS feature Electronic store	**No**	**8**	9	8			20	10	9		9
	Yes	**83**	91	92	100	100	80	90	91		91

Table (6-16). Objective 1-2 (Features of these information systems) Results.

Discussion of the first objective (Cont'd)

1-Identify the current Admission and Registration Information Systems in the Egyptian Universities concerning the following:

1-3 Functions of these information systems	Questions **12-1: 12-8**

Table (6-17) shows the results of questions 12-1 : 12-7.

Discussion

- These questions investigate the functions of the current Admission and Registration information systems.
- *Overall*: the majority of the respondents reported positively on four questions (12-1, 12-3, 12-4, and 12-6), whilst the majority reported negatively on the three remaining questions (12-2, 12-5, and 12-7).
- *The University type dimension*: This dimension has an effect on the answers to question 12-6 (*Using the historical data to describe the Students' history*) whilst it has no effect on the rest of these questions. So, whether the respondent belongs to a Private or Governement University, it does not make any difference in his response to all questions, except for 12-6. The majority of respondents in both University types reported positively (i.e. 12-1, 12-3, and 12-4) and negatively (i.e. 12-2, 12-5, and 12-7) on the same questions of this construct. The responses are ranging from 53 to 94% for the positive questions, whilst from 66 to 87% for the negative questions. On question 12-6 whilst the overall majority reported positively, 61% of the respondents from Government Universites reported negatively.
- *The respondent position dimension:* This dimension also has an effect on the answers of two questions 12-6 and 12-7 (*Using the historical data to describe the Students' history, Using external data to enhance the quality of decisions*), whilst it has no effect on the remaining questions on which either the repondent is a Dean, Associate Dean, Registrar, Admission Officer, or Other, this does not affect his response to these

questions. The majority of all positions reported positively (i.e. 12-1, 12-3, and 12-4) and negatively (i.e. 12-2, and 12-5) on the same questions of this construct. The responses are ranging from 50 to 92% for the positive questions, and from 50 to 90% for the negative questions.

- *The use of CBIS dimension:* This dimension has no effect on the answers to these questions. All respondents who answered these questions should be using CBIS (again this is because Non-CBIS users would have moved to question 13 after the branching question 3). On the positive questions the responses are ranging from 55 to 86%, whilst on the negative questions responses are ranging from 68 to 84%.
- A number of questions were designed to validate each others in the ARDSSQ. For example questions 5 and 12-6; when we try to validate the results of them together (i.e. 5 and 12-6) we found the overall majority of respondents reposted positively on both. Another example is found between questions 12-1 & 12-2 in one hand and 11-1 & 11-2 in the other hand; the overall majority reported positively on 11-1 and 12-1, whilst negatively on 11-2 and 12-2. This is another source of validity to this research questionnaire.
- The overall majority of negative responses on question 12-7 (*Using external data to enhance the quality of decisions*), reveals that external data is not used in both University types. However, Deans are the only respondents who reported positively when asked about using external data. This implies that users' information needs differ by management levels. This is also supported by the answers to question 12-5 (*Finding relationships between a student's data fields*) on which Deans are the only respondents who reported positively (50%), whilst the remaining positions reported negatively.
- Based on the previous discussion:
  - The functions of the current Admission and Registration information systems are affected by these two dimensions; the University type and the manager's position.

- General statistics is the system function that received the highest positive responses (77%). This indicates the importance of providing general statistics to decision makers.
- Different management levels require different information needs.
- The Admission and Registration information systems in the Egyptian Universities have the following functions:
    - Student description reports;
    - General statistics;
    - Classifying students into similar groups;
    - Using the historical data to describe the Students' history (*only in the Private Universities*).
- The Admission and Registration information systems in the Egyptian Universities do not have the following functions:
    - Student performance prediction;
    - Finding relationships between a student's data fields;
    - Using external data to enhance the quality of decisions.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.12-1 IS Function Reports	**No**	**20**	6	47	50	23	25	20	12		23
	Yes	**69**	94	53	50	77	75	80	88		77
Q.12-2 IS Function Prediction	**No**	**71**	87	76	60	83	90	80	85		84
	Yes	**15**	13	24	40	17	10	20	15		16
Q.12-3 IS Function Statistics	**No**	**12**	13	14	8	8	10	10	21		14
	Yes	**77**	87	86	92	92	90	90	79		86
Q.12-4 IS Function Classification	**No**	**25**	30	25	17	15	25	20	41		28
	Yes	**64**	70	75	83	85	75	80	59		72
Q.12-5 IS Function Relationships	**No**	**65**	75	74	50	82	75	80	79		76
	Yes	**22**	25	26	50	18	25	20	21		24
Q.12-6 IS Function Use History	**No**	**39**	33	61	42	54	50	33	42		45
	Yes	**48**	67	39	58	46	50	67	58		55
Q.12-7 IS Function External Data	**No**	**57**	68	66	33	92	65	75	70		68
	Yes	**28**	32	34	67	8	35	25	30		32

Table (6-17). Objective 1-3 (Functions of these information systems) Results.

6-14-2 Discussion of the second objective

2-Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:

2-1 The managers' perspectives towards the role of computers and the ideal Admission and Registration information system	Questions **13: 22**

Table (6-18) shows the results of questions 13 : 22.

Discussion

- These questions investigate the managers' perspectives towards the role of computers and the ideal Admission and Registration information system.
- *Overall*: the majority of the respondents reported positively on these nine questions 13: 15, and 17: 22, whilst the majority reported negatively on one question 16.
- *The University type dimension*: This dimension has an effect on the answers of two questions 13 and 15 (*I believe that the main role of computer is electronic data storage, The higher the rank of the decision maker is in the chain of command, the reports produced by the system are required to contain more detail*), whilst it has no effect on the rest of these questions on which whether the respondent belongs to a Private or Government University, it does not make any difference in his response to them. The majority of respondents reported positively (i.e. 13: 15 and 17: 22), and negatively only on question 16 (*The fact that my competitors-outside the organization- may have access to the same information makes it less useful*) on the same questions, except for questions 13 and 15. While the majority reported positively on both 13 and 15, the Private University respondents reported negatively on both (52 and 57% respectively). On the remaining questions, the responses range from 53 to 100% for the positive questions and from 60 to 71% for the negative question.
- *The respondent position dimension:* This dimension also has an effect on the answers of three questions 15, 18, and 19. However, it has no effect on the remaining questions

on which either the repondent is a Dean, Associate Dean, Registrar, Admission Officer, or Other, this does not affect his responses to these questions (i.e. 13, 14, 16, 17, and 20: 22). The responses to these questions are ranging from 50 to 100% for the positive questions (i.e. 13: 14, 17, and 20: 22), whilst from 54 to 89% for the negative question (i.e. 16). On question 15, 18, and 19 the majority of respondents reported positively, however, negatively by: 60 % of the Admission Officers on 15, 55% of the Deans on 18, and 56% of the Admission Officers on 19.

- *The use of CBIS dimension:* This dimension has no effect on the answers to these questions. Either repondents use CBIS or not, this does not affect their responses. The majority of respondents (CBIS users or non-users) reported positively on the same nine questions of this construct (i.e. 13: 15, and 17: 22), wihlst negatively on the same question (i.e. 16). On the positive questions the responses are ranging from 53 to 100%, whilst on the negative question from 65 to 69%.
- The only question on which the overall majority reported negatively was 16 (*The fact that my competitors-outside the organization- may have access to the same information makes it less useful*), and this fact reflects a good understanding of the use of information by the respondents in both University types for all the managers' levels.
- The overall majority of respondents reported positively on question 13 (*I believe that the main role of computer is electronic data storage*), however this majority came from the Government Universities (80%), whereas the Private Universities reported negatively (52%). This finding reflects a better understanding to the role of computers in the Private Universities, and at the same time a narrower scope to this role in the Government Universities.
- On questions 14 and 21 (*I believe that one of the computer roles is to be a decision maker, The Admission and Registration information system should be able to help managers take decisions*) the overall majority reported positively (60%, 99% respectively) regardless of their University type, Position, and use of CBIS. Based on

this finding we can say that there is a need for a DSS in the area of Admission and Registration function in the Egyptian Universities.

- On question 15 (*The higher the rank of the decision maker is in the chain of command, the reports produced by the system are required to contain more detail*), the majority of respondents from the Government Universities reported positively (79%), whilst negatively from Private Universities (57%). This finding enhances the thought that the managers' computer awareness in Private Universities' is higher than those in Government Universities' because the later don not understand the relationship between the level of report details and the manager's level in the chain of command.
- The results of question 21 (*The Admission and Registration information system should be able to help managers take decisions*) showed that 98% of the Private Universities, and 100% of the Government Universities think that their Admission and Registration information system should be able to help managers take decisions. This means that there is a need for DSS to be developed for the Admission and Registration functions.
- Based on the previous discussion:
  - The managers' perspectives towards the role of computers and the ideal Admission and Registration information system are affected by these two dimensions; the University type and the manager's position.
  - The use of CBIS doens not affect the managers' perspectives towards the role of computers and the ideal Admission and Registration information system.
  - Respondents from the Private Universities have a better understanding to the managers' perspectives towards the role of computers and the ideal Admission and Registration information system.
  - There is a need for DSS to be developed for the Admission and Registration functions.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.13 PC role is Data store	**No**	**52**	52	20	19	23	21	50	50	21	41
	Yes	**111**	48	80	81	77	79	50	50	79	59
Q.14 PC role Decision Maker	**No**	**66**	35	44	46	40	47	30	34	47	35
	Yes	**97**	65	56	54	60	53	70	66	53	65
Q.15 Higher Rank Higher Details	**No**	**56**	57	21	31	20	23	60	50	22	44
	Yes	**107**	43	79	69	80	77	40	50	78	56
Q.16 Same Info Less Useful	**No**	**104**	60	71	79	72	54	89	65	69	65
	Yes	**52**	40	29	21	28	46	11	35	31	35
Q.17 IS Internal Data Sources	**No**	**19**	19	8		10	**9**		24	10	13
	Yes	**142**	81	92	100	90	91	100	76	90	87
Q.18 IS external Data	**No**	**64**	47	38	55	36	40	30	42	33	47
	Yes	**91**	53	62	45	64	60	70	58	67	53
Q.19 IS should forecast	**No**	**18**	21	6	4	7	13	56	8	4	17
	Yes	**137**	79	94	96	93	87	44	92	96	83
Q.20 IS should improve Decision Quality	**No**	**9**	12	2		3	2	30	8	3	8
	Yes	**153**	88	98	100	97	98	70	92	97	92
Q.21 IS help managers take decisions	**No**	**1**	2						2	1	
	Yes	**161**	98	100	100	100	100	100	98	99	100
Q.22 More Actionable More Acceptable	**No**	**3**	3	1			2		4	3	1
	Yes	**158**	97	99	100	100	98	100	96	97	99

Table (6-18). Objective 2-1 (The managers' perspectives towards the role of computers and the ideal Admission and Registration information system) Results.

Discussion of the second objective (Cont'd)

2-Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the foclllowing:

2-2 The decisions that this DSS is expected to take	Questions **23-1: 23-23**

Table (6-19) shows the results of questions 23-1 : 23-18.

Discussion

- These questions investigate the decisions that the Admission and Registration DSS is expected to take.
- The questions of this construct are very important because based on their results the proposed Admission and Registration DSS is going to be developed. For the development of the proposed Admission and Registration DSS, all the decisions on which the majority of respondents report positively will be implemented. On the other hand, when respondents report negatively on any question(s); that/those will not be part of the implementation process. Moreover, if the proposed DSS is to be implemented in a specific University adjustments (i.e. adding or removing decisions) need to happen to reflect the University-specific information needs and environment.
- *Overall*: the majority of the respondents reported positively on 16 questions, whilst negatively only on two question 23-2 and 23-7.
- *The University type dimension*: This dimension affects the responses of four questions (23-7, and 23-9: 23-11), on 23-7 the majority reported negatively whilst the Private Universities reported positively (also the majority of the CBIS users reported this question negatively), on 23-9: 23-11 the majority reported positively whilst the Private Universities reported negatively (their experiences with their current systems might justify why they reported negatively). This dimension has no effect on the remaining fourteen questions, that is either the respondent belongs to Private or Public University his answers to these questions are not affected. On the positive questions responses are

ranging from 56 to 89%, on the negative questions responses are ranging from 66 to 74%.

- *The respondent position dimension:* This dimension affects the responses of nine questions (23-5, 23-7, 23-9: 23-11, 23-13, and 23-16: 23-18), the majority reported negatively on 23-7, whilst 53% Registrars reported positively, also the majority reported positively on the remaing eight (23-5, 23-9: 23-11, 23-13, and 23-16: 23-18) whilst: 60% Admission Officers, 70% Admission Officers, 67% Admission Officers, 60% Registrars, 56% Admission Officers, 60% Admission Officers, 56% Admission Officers, and 59% Registrars reported negatively on these questions respectively. This dimension does not affect the responses of the remaining nine questions on which the responses are ranging from 50 to 97% for the positive questions, and from 57 to 80% for the negative questions.
- *The use of CBIS dimension:* This dimension affects the responses of three questions (23-7, 23-10, and 23-11), on 23-7 the majority reported negatively whilst the CBIS users reported positively (also the majority of the Private Universities reported this question negatively), on 23-10 and 23-11 the majority reported positively whilst the CBIS users reported negatively (their experiences with their current systems might justify why they reported negatively). This dimension has no effect on the remaining fifteen questions, that is either the respondent is CBIS user or not his answers to these questions are not affected. On the positive questions the responses are ranging from 53 to 88%, whilst on the negative questions from 55 to 76%.
- The only two questions on which the majority reported negatively were 23-2 and 23-7 (*Provide unconditional offer for new applicant, Hold the applicant until the following term/year*). Concerning question 23-2, this decision is very critical and not many managers would be able to afford the consequences of mistakes of this type (i.e. providing unconditional offer would cost any Academic Institution lots of money). And concerning question 23-7, it was reported negatively from the respondents that belong

to Government Universities where most of which still do not have this mechanism (i.e. holding applicants for the coming term/year). Based upon this, 23-2 will not be implemented in the proposed DSS, however, 23-7 will be implemented as the percentage of respondents who said No (50.3%) is very close to those who said Yes (49.7%).

- Based on the previous discussion:
  - The decisions that the Admission and Registration DSS is expected to take are affected by these dimensions; the University type, the manager's position, and the use of CBIS.
  - The Admission and Registration DSS should be able to take the following decisions:
    - Accept or reject a new applicant;
    - Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records;
    - Predict the new applicants that will join the faculty/college/institute this term/year based on government statistics on secondary school students;
    - Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records besides other records like the government statistics;
    - Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major;
    - Hold the applicant until the following term/year (only for Private Universities);
    - Accept or reject the applicant who is transferred from another educational institution;
    - Accept or reject the applicant who is transferred from another educational institution based on our transfer history records (only for Government Universities);

- Predict a student's performance based on the students' history we keep (only for Government Universities);
- Predict a course's results based on the courses' history we keep (only for Government Universities);
- Classifying students into similar groups;
- Predict a student's performance based on the group that he/she belongs to;
- Set the student status to "On probation";
- Predict the "On probation" students based on the students' history we keep;
- Make relationships between students' performance and academic departments;
- Forecast course booking;
- Decide on Student abandonment.

- The Admission and Registration DSS should not take the following decision:
  - Provide unconditional offer for new applicant.
- It is recommended in future use of this questionnaire that questions 23-21: 23-23 would be eliminated as they did not receive any response at all.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.23-1 Accept Reject New Applicant	No	45	28	30	23	31	41	40	17	31	27
	Yes	111	72	70	77	69	59	60	83	69	73
Q.23-2 Unconditional Offer	No	106	74	66	58	57	75	80	74	76	64
	Yes	47	26	34	42	43	25	20	26	24	36
Q.23-3 Predict New Applicant by History	No	41	30	24	40	24	20	33	26	18	33
	Yes	113	70	76	60	76	80	67	74	82	67
Q.23-4 Predict new Applicant by Statistics	No	47	33	29	44	31	20	50	27	22	36
	Yes	109	67	71	56	69	80	50	73	78	64
Q.23-5 Predict new Applicant by History and Statistics	No	38	31	20	36	10	27	60	17	15	31
	Yes	118	69	80	64	90	73	40	83	85	69
Q.23-6 Applicant Major Match	No	35	23	22	8	17	30	30	25	27	19
	Yes	120	77	78	92	83	70	70	75	73	81
Q.23-7 Hold Applicant	No	77	48	52	52	55	47	60	48	55	47
	Yes	76	52	48	48	45	53	40	52	45	53
Q.23-8 Accept Reject Transfer	No	47	38	26	25	31	35	40	28	31	31
	Yes	106	62	74	75	69	65	60	72	69	69
Q.23-9 Accept Reject Transfer by History	No	67	57	36	29	43	47	70	45	40	47
	Yes	85	43	64	71	57	53	30	55	60	53
Q.23-10 Predict Performance History	No	70	53	42	35	41	51	67	46	38	52
	Yes	82	47	58	65	59	49	33	54	62	48
Q.23-11 Predict Course Results	No	69	53	41	35	38	60	44	42	36	52
	Yes	83	47	59	65	62	40	56	58	64	48
Q.23-12 Classify Students into Groups	No	19	11	14	17	3	17	11	13	14	12
	Yes	132	89	86	83	97	83	89	88	86	88
Q.23-13 Predict Student Performance by Group	No	48	37	29	26	24	43	56	25	28	35
	Yes	103	63	71	74	76	57	44	75	72	65
Q.23-14 Set to On Probation	No	39	13	36	14	41	43	25	11	35	21
	Yes	105	87	64	86	59	57	75	89	65	79
Q.23-15 Predict On Probation	No	46	23	36	9	48	49	22	17	32	30

	Yes	**102**	77	64	91	52	51	78	83	68	70
Q.23-16 Relate Performance to department	**No**	**42**	43	18	9	17	28	60	35	17	36
	Yes	**110**	57	82	91	83	72	40	65	83	64
Q.23-17 Forecast Course Booking	**No**	**40**	28	26	26	28	36	56	13	20	31
	Yes	**111**	72	74	74	72	64	44	88	80	69
Q.23-18 Student Abandonment	**No**	**58**	44	36	17	38	59	38	33	40	38
	Yes	**91**	56	64	83	62	41	63	67	60	62

Table (6-19). Objective 2-2 (The decisions that this DSS is expected to take) Results.

Discussion of the second objective (Cont'd)

2-Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:

2-3 DSS functions	Questions **24-1: 24-11**

Table (6-20) shows the results of questions 25-1 : 25-9.

Discussion

- These questions investigate the ideal Admission and Registration information system's functions.
- *Overall*: the majority of the respondents reported positively on all questions.
- *The University type dimension*: This dimension has an effect on the answers to the first question 24-1 only. However, has no effect in the remaining questions 24-2: 24-9. With regard to question 24-1, 53% of respondent from the Private Universities reported negatively, whilst 62% of respondent from the Government Universities reported positively. On the remaining questions, this dimension has no effect on the answers. That is, whether the respondent belongs to a Private or Government University, it does not make any difference in his response to questions 24-2: 24-9 on which the majority of respondents in both University types reported positively ranging from 66 to 99%.
- *The respondent position dimension:* Again, this dimension has an effect on the answers to the first question 24-1 only. However, has no effect in the remaining questions 24-2: 24-9. With regard to question 24-1: 52% of Registrars and 52% of the Others reported negatively, whilst: 76% of Deans, 68% of Associate Deans, and 56% of Admission Officers reported positively. On the remaining questions of this construct, this dimension has no effect on the answers. That is, whatever the respondent position is, it does not make any difference in his response to questions 24-2: 24-9 on which the majority of respondents in all positions reported positively. The responses are ranging from 51 to 100%.

- *The use of CBIS dimension:* This dimension also has no effect on the answers to these questions. Either the repondent uses a CBIS or not, this does not affect the response. The majority of all respondents (CBIS users or non-users) reported positively on all the questions, the responses are ranging from 53 to 97%.
- The results of these questions 24-2: 24-9 are very logical and consistent.
- The least important function is *predicting new applicants' performance* (Q.24-1) however reported positively in the overall responses, it was negatively reported by the majority of the respondents from Private Universities (53%), and the positions that rejeceted the use of this function are *Registrars and Others*. On the other hand, *predicting current students' performance* (24-2) received positive responses from both University types, all positions, either from users or non-users of CBIS.
- The results of the question *using historical data* (24-6), which received positive responses from all respondents regardless of their University types, positions, CBIS users or non-users, fosters the need for a DW in the Admission and Registration information Systems.
- Based on the previous discussion:
  - The ideal Admission and Registration information system functions are affected by both the University type and the manager's position dimensions.
  - The ideal Admission and Registration information system functions are NOT affected by the use of CBIS dimension.
  - The ideal Admission and Registration information system should have the following functions:
    - Predict new applicants' performance (*only for the Government Universities*);
    - Predict current students' performance;
    - Producing student description reports;
    - Provide general statistics;
    - Student classification into groups;

- Using historical data;
- Being able to use external data;
- Finding relationships between students' data fields;
- Gives the user the abilty to create ad hoc reports.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.24-1 IS Function New Applicant Performance Prediction	No	**68**	53	38	24	32	52	44	52	40	47
	Yes	**88**	47	62	76	68	48	56	48	60	53
Q.24-2 IS Function Current Student Prediction	No	**39**	31	22	12	14	41	33	23	29	23
	Yes	**115**	69	78	88	86	59	67	77	71	77
Q.24-3 IS Function Reports	No	**10**	2	9	4	3	15		2	12	2
	Yes	**148**	98	91	96	97	85	100	98	88	98
Q.24-4 IS Function Statistics	No	**5**	7	1		3			8	3	3
	Yes	**154**	93	99	100	97	100	100	92	97	97
Q.24-5 IS Function Classify Students to Groups	No	**24**	20	12	8	7	13		30	12	18
	Yes	**135**	80	88	92	93	87	100	70	88	82
Q.24-6 IS Function Use Historical Data	No	**13**	5	10	4	10	11		9	13	4
	Yes	**144**	95	90	96	90	89	100	91	87	96
Q.24-7 IS Function Use External Data	No	**47**	34	27	28	17	35	40	32	33	27
	Yes	**111**	66	73	72	83	65	60	68	67	73
Q.24-8 IS Function Relationships	No	**22**	15	13	4	13	17	40	11	16	12
	Yes	**137**	85	87	96	87	83	60	89	84	88
Q.24-9 IS Function Ad Hoc Reports	No	**51**	25	39	38	28	49	25	23	41	29
	Yes	**100**	75	61	63	72	51	75	77	59	71

Table (6-20). Objective 2-3 (DSS functions) Results.

Discussion of the second objective (Cont'd)

2-Extract the information requirements for a new Admission and Registration DSS in the Egyptian Universities concerning the following:

2-4 DSS characteristics	Questions **25-1: 25-11**

Table (6-21) shows the results of questions 25-1 : 25-9.

Discussion

- These questions were designed to investigate the ideal Admission and Registration information system's characteristics.
- *Overall*: the majority of the respondents reported positively on all questions.
- *The University type dimension*: This dimension has no effect on the answers to these questions. So, whether the respondent belongs to a Private or Governement University, it does not make any difference in his response. The responses are ranging from 83 to 99%.
- *The respondent position dimension:* This dimension also has no effect on the answers to these questions. Similar responses were obtained to the system requirements regardless of the respondent position. The majority of all positions reported positively on all the questions, the responses are ranging from 78 to 100%.
- *The use of CBIS dimension:* This dimension also has no effect on the answers to these questions. Either the repondent is a CBIS user or non-user, this does not affect his responses. The majority of all respondents (users or non-users) reported positively on all the questions, the responses are ranging from 76 to 100%.
- The results of these questions (25-1: 25-9) are very logical and consistent.
- Based on the previous discussion:
  - The ideal Admission and Registration information system characteristics are NOT affected by any of these dimensions; the University type, the manager's position, or the use of CBIS.

- The ideal Admission and Registration information system should have the following characteristics:
  - Easy to use;
  - Requires minimum training;
  - User involvement in design of the system;
  - Able to grow;
  - Flexible;
  - Integrated;
  - Have E-mail facility;
  - Web-accessible;
  - And cost effective.
- Q.25-11 should be removed in future use of this questionnaire. No responses were received on it.

Questions		Overall	University Type		Respondent Position					Use CBIS?	
			Private	Government	Dean	Associate Dean	Registrar	Admission Officer	Other	No	Yes
			%	%	%	%	%	%	%	%	%
Q.25-1 IS Characteristic Ease	**No**	7	10	1			2		13		8
	Yes	**148**	90	99	100	100	98	100	87	100	92
Q.25-2 IS Characteristic Min Training	**No**	**13**	2	13	20	10	9		2	16	2
	Yes	**143**	98	88	80	90	91	100	98	84	98
Q.25-3 IS Characteristic User-Involvement	**No**	**24**	14	17	12	17	16	22	15	24	9
	Yes	**130**	86	83	88	83	84	78	85	76	91
Q.25-4 IS Characteristic Grow	**No**	**8**	10	2	4		2		13	3	7
	Yes	**148**	90	98	96	100	98	100	87	97	93
Q.25-5 IS Characteristic Flexible	**No**	7	8	2	4		2		11	2	7
	Yes	**148**	92	98	96	100	98	100	89	98	93
Q.25-6 IS Characteristic Integrated	**No**	**8**	5	5	8	7	2		7	8	3
	Yes	**146**	95	95	92	93	98	100	93	92	97
Q.25-7 IS Characteristic E-mail	**No**	**16**	8	12	8	3	14	10	13	15	7
	Yes	**139**	92	88	92	97	86	90	87	85	93
Q.25-8 IS Characteristic Accessible through WEB	**No**	**16**	7	13	17	10	11	10	7	14	8
	Yes	**139**	93	87	83	90	89	90	93	86	92
Q.25-9 IS Characteristic Cost-effective	**No**	**20**	12	14	16	7	20		11	16	10
	Yes	**136**	88	86	84	93	80	100	89	84	90

Table (6-21). Objective 2-4 (DSS characteristics) Results.

6-15 Analysis of open-ended questions

Originally the questionnaire includes five questions (**11**, **12**, **23**, **24**, and **25**) where respondents had extra space to add any other function, feature or decision where it is not listed in the questionnaire. Only the space provided by questions **11**, **23**, **24**, and **25** received comments and suggestions from **10** respondents, whilst space provided by question **12** did not receive any comments or suggestions. The following table (6-22) describes these issues.

Respondent	University type	Questions	Comment or Suggestion
Associate Dean	Government	**11** **23**	-Students' graduation rules -Student-Major match -Abandonment rules
Dean	Government	**25**	-The system should contain self-training kit to help new users
Registrar	Government	**23**	-The system should be able to apply the mercy rules
Dean	Government	**25**	-Perfect security system
Registrar	Private	**11** **23**	-Print student's reports -Finding the courses that each student should complete before graduation
Registrar	Private	**24**	-Class scheduling capacity -Students' direct interaction with the system
Associate Dean	Government	**23** **24** **25**	-Organising job fair -Graduates' profiles -Applying Admission rules -Student training programmes
Registrar	Government	**23**	-Applying absence ratio, if exceeded then automatically drop the course
Registrar	Private	**11** **23**	-Statistics on Students' data. -Warning to the candidate "On

			Probation" students.
Associate Dean	Government	**23**	-Ordered list of students' GPA -List of students with Honor degree

Table (6-22). Questionnaires with suggestions.

Note:

1. The 10 respondents were distributed as 7 from government Universities and 3 from private. These are 2 Deans, 3 Associate Deans, and 5 Registrars.
2. Open-ended questions 11, 23, 24, and 25 were completed by the respondents and hence it was valid to include them in the questionnaire.
3. Suggestions can be classified into three categories:

   1-Those which are already included in the questionnaire. Example applying Admission rules;

   2-Those which are irrelevant to the topic of the questionnaire. Example organising job fair;

   3-Some relevant suggestions. Example the system should be able to apply the mercy rules.

6-16 Representing the Objectives/Constructs

As mentioned earlier each of the ARDSSQ objective/construct represents a group of questions that are supposed to measure this ojbective/construct. For example questions 1:10 are measuring the first construct, 11-1:11-6 the seconds..etc.

In the next sections Chi Square and Canonical Correlation will be used as analysis techniques. To perform the analysis techniques, questions that are supposed to measure a certain ojbective/construct were summed up together to represent them in the analysis (Keil, et al., 2000). Then the questions are summed and divided by their number to give the average. In Chi Square no differences were obtained while running the technique based on either summations or averages. However, the results listed in the following sections are based on the questions' summations.

6-17 Generalizations about the population

It is worth mentioning here that it is not the intention of this research to measure the strength and/or direction of any possible relationships- if found- however, the intention is just to draw some generalizations about the population which will be helpful in three directions:

1. Better understanding for the environment under study;
2. Exploring possible relationships that may be useful for future research;
3. This could also explain some of the differences in responses. For example the response of certain questions might be affected by the respondent position (e.g. do you have a PC at your desk?) in some questions or by the University type (do you use a CBIS?).

The reason why this research is not targeting these relationships in detail is because the objective of this research is to define the current Admission and Registration systems in Egyptian Universities and to identify the information needs for a new system, then building

this new required system. Since the development of the new system is not affected by such relationships this is why this research is not exploring them in detail.

Table (6-23) contains some of the possible relationships that are expected to exist and would be helpful for future research and could also explain some of the responses diversity.

The relationships that will be investigated
1-There is no relationship between the University type and the use of CBIS; 2-There is no relationship between the Respondent position and the use of CBIS; 3-There is no relationship between the Respondent position and the acceptance to role of computers as data stores; 4-There is no relationship between the respondent position and the acceptance to role of computers as decision makers; 5-There is no relationship between the Respondent position and the availability of PC's at their desks.

Table (6-23). The assumed relationships

In the following section these relationships will be examined using the *non-parametric* test Chi Square and Canonical correlation analysis.

6-17-1 The Chi Square Test

Based on the type of variables (dichotomous) and the relationship that would be explored, the *non-parametric* test Chi Square will be used. Tests concerned with nominal or ordinal levels of measurement are called non-parametric tests or distribution free tests. Distribution free means that these tests are free of assumptions regarding the data distribution of the parent population. The Chi-Square Test is one of these tests (Zikmund, 2000; Mason, et al., 1999; Bee and Bee, 1990). The Chi-Square Test procedure tabulates a variable into categories and computes a chi-square statistic.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where χ^2 =chi-square statistic, O_i=observed frequency in the *i*th cell, E_i=expected frequency in the *i*th cell, *k* is the number of categories.

This test compares the observed and expected frequencies in each category to test either that all categories contain the same proportion of values or that each category contains a user-specified proportion of values (Zikmund, 2000). Cooper and Schindler (1998: 482) said "Probably the most widely used nonparametric test of significance is the chi-square test".

Data and assumptions. Data could be ordered or unordered numeric categorical variables (ordinal or nominal levels of measurement). Nonparametric tests do not require assumptions about the shape of the underlying distribution. The data is assumed to be obtained by using a random sample.

Bryman and Cramer (1999: 124) said that there is a restriction on the use of the chi square test "When the expected frequencies are small with only two categories (one degree of freedom), the number of cases expected to fall in these categories should be at least 5 before this test can be applied".

Discussion of the Chi Square test results

1-There is no relationship between the University type and the use of CBIS. That is the proportion of private Universities who are using computerized Admission and Registration information systems does not differ from that proportion in the government ones.

		University Type	
		Private %	Government %
Use of CBIS	No	8	65
	Yes	92	35
	Total	N= 61	N= 106
		$X^2 = 50.79$	S, $p < .05$

Table (6-24). University type and CBIS use.

Table (6-24) indicates that there is a difference in the observed percentages between the proportions of Government and Private Universities regarding whether or not they are using CBIS. The table also indicates that this difference is significant ($p < .05$). This means that we would reject the assumption that there is no relationship between the University type and the use of CBIS. Alternatively, there is a significant relationship between the University type and the use of CBIS by which the proportion of private Universities who are using CBIS differs from that proportion in the government ones.

2-There is no relationship between the Respondent position and the use of CBIS. That is the proportion of Deans who are using computerized Admission and Registration information systems does not differ from that proportion as for the Associate Deans, Registrars, Admission Officers, and Others.

		Respondent position				
		Dean %	Associate Dean %	Registrar %	Admission Officer %	Others %
Use of CBIS	No	56	53	58	-	29
	Yes	44	47	42	100	71
	Total	N= 27	N= 30	N= 48	N= 10	N = 52
			$X^2 = 19.19$		S, $p < .05$	

Table (6-25). Respondent position and CBIS use.

Table (6-25) indicates that there is a difference in the observed percentages between the different respondent positions and their use of a CBIS. It also indicates that this difference is significant ($p < .05$). This means that we would reject the assumption that there is no relationship between the Respondent positions and the use of CBIS. Alternatively, there is a significant relationship between the respondent position and the use of CBIS by which the proportion of Deans who are using CBIS differs from that proportion with regard to the Associate Deans, Registrars, Admission Officers, and Others.

3-There is no relationship between the Respondent position and the acceptance to role of computers as data stores. That is the proportion of Deans who believe that being a data store is the main role of computers does not differ from that proportion as for the Associate Deans, Registrars, Admission Officers, and Others.

		Respondent position				
		Dean %	Associate Dean %	Registrar %	Admission Officer %	Others %
Computers as Data Stores	No	19	24	21	50	48
	Yes	91	76	79	50	52
	Total	N= 27	N= 30	N= 48	N= 10	N = 52
			$X^2 = 14.42$		S, $p < .05$	

Table (6-26). Respondent position and the data store role.

Table (6-26) indicates that there is a difference in the observed percentages between the different respondent positions and their acceptance to the role of computers as data stores. It also indicates that this difference is significant ($p < .05$). This means that we would reject the assumption that there is no relationship between the Respondent positions and their acceptance to the role of computers as data stores. However, this is true for the positions of Deans, Associate Deans, Registrars, and Others, but as for the Admission Officers 50 % believe No, and 50 % Yes, so they are in-different to this role. Alternatively, there is a significant relationship between the respondent position and acceptance to role of computers as data stores by which the proportion of Deans who believe that being a data store is the main role of computers differs from that proportion as for the Associate Deans, Registrars, and Others.

4-There is no relationship between the respondent position and the acceptance to role of computers as decision makers. That is the proportion of Deans who believe that being a decision maker is the main role of computers does not differ from that proportion as for the Associate Deans, Registrars, Admission Officers, and Others.

		Respondent position				
		Dean %	Associate Dean %	Registrar %	Admission Officer %	Others %
Computers as decision makers	No	45	40	46	30	34
	Yes	55	60	54	70	66
	Total	N= 27	N= 30	N= 48	N= 10	N = 52
			$X^2 = 2.45$		NS, p > .05	

Table (6-27). Respondent position and the decision maker role.

Table (6-27) indicates that there is a difference in the observed percentages between the different respondent positions and their acceptance to the role of computers as decision makers. However, the table also indicates that this difference is not significant ($p > .05$). This means that we would not reject the assumption that there is no relationship between the respondent positions and the role of computers as being decision makers.

5-There is no relationship between the Respondent position and the availability of PC's at their desks. That is the proportion of Deans who have PC's on disks does not differ from that proportion as for the Associate Deans, Registrars, Admission Officers, and Others.

		Respondent position				
		Dean %	Associate Dean %	Registrar %	Admission Officer %	Others %
Having a PC on desk	No	16	10	35	10	46
	Yes	84	90	65	90	54
	Total	N= 27	N= 30	N= 48	N= 10	N = 52
			$X^2 = 22.77$		S, p < .05	

Table (6-28). Respondent position and PC availability.

Table (6-28) indicates that there is a difference in the observed percentages between the different respondent positions and their ownership of a PC on their desks. It also indicates that this difference is significant ($p < .05$). This means that we would not accept the assumption that there is no relationship between the respondent positions and the ownership of a PC on desk. Alternatively we can say that there is a significant relationship

between the respondent position and the ownership of a PC on his desk by which the proportion of Deans who have a PC's on desks differ from that proportion for the Associate Deans, Registrars, Admission Officers (Admission Officer and Associate Deans are the same), and Others.

6-17-2 The Canonical Correlation analysis

Canonical correlation is one of the multivariate analysis techniques. It is used to study the interrelationships between two multiple variable sets; one set represents the independents (predictor variables) and the other set represents the dependents (criterion variables). This analysis can handle both variable types; the quantitative (metric) and the qualitative (nonmetric) (Cheng, 1995; Cliff, 1987). Hair et al., (1995: 327) said "Canonical correlation analysis is the only multivariate technique that can handle the nonmetric dependent and nonmetric independents". Other statistical techniques (i.e. MANOVA) can handle nonmetric independents but requires the dependents to be metric. The general form of canonical analysis is (Walker, 2001; Hair, et al., 1995):

$$b_1 Y_1 + b_2 Y_2 + \ldots + b_n Y_n = a_1 X_1 + a_2 X_2 + \ldots a_n X_n$$

Where Y's represent the dependent variables (one canonical variate) and the X's represent the independent variables (the other canonical variate). Canonical analysis does not impose many restrictions on the data type. Hair et al. (1995: 329) said, "Canonical correlation is the most generalized member of the family of multivariate statistical techniques". The objective of Canonical analysis is to decide whether the two variable sets are related or not and the magnitude of their relationship (Walker, 2001). Canonical analysis starts with some Canonical functions that consist of two variates (one for the independents and one for the dependents), the number of possible functions equals the number of variables in the smallest variate set. The first canonical function found should account for the maximum amount of relationship between the two sets of variables. The strength of the relationship is reflected by the canonical correlation. The function that should be chosen should be

significant and has the highest magnitude (Cliff, 1987).

- *The reasons of using of Canonical correlation analysis are*:

1-The type of data used in this research (nonmetric/qualitative);

2-The emphasis is to study the effect of set of variables on another (i.e. not the effect of every single independent on the dependents, otherwise other techniques could have been used like ANOVA-given that the data meets the technique requirements).

- *The Canonical correlation analysis results:*

The canonical correlation analysis will be used to express the strength of the relationship between these two sets of variables: the three dimensions of the analysis (University type as Y_1, Manager's position as Y_2, and CBIS use as X_3) as independent variables and the ARDSSQ seven dimensions as dependent variables (Construct_1: Construct_7). The following table (6-29) illustrates the results of applying Canonical correlation analysis to the prestated two variable sets.

Canonical analysis summary

- Variables in set 1:
Y_1 (University type), Y_2 (Manager's position), X_3 (CBIS use)
- Variables in set 2:
Construct_1, Construct_2, Construct_3, Construct_4, Construct_5, Construct_6, Construct_7 (Refer to table 6-1)
- Number of complete cases: 167

Section 1- Canonical Correlations

Function Number	Eigenvalue	Canonical Correlation	Wilks Lambda	Chi-Square	D.F.	P- Value
1	0.932	0.965	0.0591295	453.898	21	0.0000
2	0.086	0.294	0.878908	20.7166	12	0.0547
3	0.037	0.193	0.962535	6.12867	5	0.2939

Section 2- Coefficients for Canonical Variables of the First Set

Y_1	-0.035	1.322	0.636
Y_2	0.012	-1.103	0.613
X_3	-1.016	0.440	0.485

Section 3- Coefficients for Canonical Variables of the Second Set

Construct_1	0.174	-0.233	0.822
Construct_2	0.752	1.205	-0.190
Construct_3	0.105	-1.118	-0.626
Construct_4	0.018	0.172	0.919
Construct_5	-0.009	-0.097	-0.477
Construct_6	-0.035	-0.213	0.710
Construct_7	0.0042	1.060	-0.507

Table (6-29). Canonical analysis results.

-The Canonical correlation results illustrated in table (6-29) tells us the following:

1. The sample *size* of this study is 167 respondents, which gives 55-to-1 ratio of observations per independent variable exceeding the 10-to-1 requirement set by some authers (Hair, et al., 1995), and others raised this requirement to 20-to-1 (Thomas, 2001);
2. Concerning the *normality and linearity*, Hair et al. (1995) mentioned that normality is not required but preferred for Canonical analysis. Bryman and Cramer (1999: 117) said "the need to meet the conditions for using parametric tests has been strongly questioned";
3. In section-1 of table (6-29) three Canonical functions have been found. The first Canonical function (number one in section-1) is the one that will be used to elaborate the relationships between the two variable sets because it is the only significant one ($P < 0.05$; statistically significant correlation at 95 % confidence level). The other two functions are not significant ($P > 0.05$), so will be relaxed here. The first function also has the highest magnitude (0.965) between the two sets of variables;
4. The chosen function's Eigenvalue is 0.932, which represents the amount of shared variance in the dependent variables that is accounted for by the independent variables.
5. The Canonical function is interpreted as follows: (Canonical function is derived from section-2 and section-3):

$$-0.035 * Y\_1 + 0.012 * Y\_2 - 1.016 * X\_3 = 0.174 * construct\_1 + 0.752 * construct\_2 + 0.105 * construct\_3 + 0.018 * construct\_4 - 0.009 * construct\_5 - 0.035 * construct\_6 + 0.004 * construct\_7$$

Where the magnitude of the variable represents its contribution to the variate it belongs to. Variables of opposite signs represent inverse relationships to each other's. That is, among the independents' variate the CBIS use (X_3) accounts for the highest effect and works on the same direction as the University type (Y_1) and both are opposite to the Manager's position (Y_2) which has the least effect on the variate. Also the second, first and third

constructs (in order) have the highest effect on the variate of the dependent variables. Among the dependent variables only the fifth and sixth constructs move in the opposite direction to the remaining constructs. The reason why the first three constructs have the highest relationship magnitudes is because they representing the current Managers' perspective towards CBIS, the current Admission and Registration IS features, and the current Admission and Registration IS functions which are highly affected by the three independents, whilst the remaining constructs are about ideal Managers' perspectives and ideal DSS decisions, functions, and characteristics where the three dimensions have little impact.

6-18 The ARDSSQ limitations

The extent to which the ARDSSQ can be used is restricted to the following:

1. It is an industry specific questionnaire. This means that it is only applicable to the higher education institutions.
2. To the Egyptian Universities, however its use can be extended to other countries providing some modifications to reflect the country-specific education system and regulations.
3. It is only relevant for evaluating the Admission and Registration IS, not any other IS.
4. The research findings apply only to the Admission and Registration IS that were running in the period of data collection (1999 – 2000).
5. The ARDSSQ that would be developed is based on the information needs of certain managers' positions to which this system is designed. These positions investigated are; Deans, Associate Deans, Registrars, Admission Officers, and Others whose positions enable them to take Admission and Registration-related decisions.

There is still a lot of additional work required in the area of evaluating information systems in general, as well as in the are of extracting user requirements. The ARDSSQ provided

potentially useful approach in evaluating Admission and Registration IS in Universities and extracting the user needs for a new ideal Admission and Registration DSS. It is the author's wish that both researchers and practitioners will find the ARDSS useful questionnaire and will use, test, and improve it to address further areas of IS.

Chapter summary

- The Admission and Registration DSS should be able to take the following decisions:
  - Accept or reject a new applicant
  - Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records
  - Predict the new applicants that will join the faculty/college/institute this term/year based on government statistics on secondary school students
  - Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records besides other records like the government statistics
  - Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major
  - Hold the applicant until the following term/year
  - Accept or reject the applicant who is transferred from another educational institution
  - Accept or reject the applicant who is transferred from another educational institution based on our transfer history records
  - Predict a student's performance based on the students' history we keep
  - Predict a course's results based on the courses' history we keep
  - Classifying students into similar groups
  - Predict a student's performance based on the group that he/she belongs to
  - Set the student status to "On probation"
  - Predict the "On probation" students based on the students' history we keep
  - Make relationships between students' performance and academic departments
  - Forecast course booking
  - Decide on Student abandonment
- The ideal Admission and Registration information system should have the following functions- according to the responses:
  - Predict new applicants' performance
  - Predict current students' performance
  - Producing student description reports

  - Provide general statistics
  - Student classification into groups
  - Using historical data
  - Being able to use external data
  - Finding relationships between students' data fields
  - Gives the user the abilty to create ad hoc reports.
- The ideal Admission and Registration information system should have the following characteristics- according to the responses:
  - Easy to use
  - Requires minimum training
  - User involvement in design of the system
  - Able to grow
  - Flexible
  - Integrated
  - Has E-mail facility
  - Wes-accessible
  - And cost effective.
- There is a significant relationship between the University type and the use of CBIS.
- There is a significant relationship between the Respondent position and the use of CBIS.
- There is a significant relationship between the Respondent position and the acceptance to role of computers as data stores.
- There is a significant relationship between the Respondent position and the availability of PC's at their desks.
- There is no relationship between the Respondent positions and the role of computers as being decision makers.
- The next chapter (seven) will discuss the ARDSS software development.

Chapter seven

The proposed Admission and Registration DSS

This chapter covers the development details of the proposed Admission and Registration DSS (ARDSS). The chapter will start with the ARDSS development methodology. The proposed DSS methodology that was introduced in chapter five will be adopted in the development of the ARDSS. The methodology consists of four modules; module 0 to identify the users' needs, module 1 to build the data warehouse, in module 2 the KDD process is applied to 1800 sample data records, and in module 3 the DSS is being developed. The discovered knowledge will be analyzed in modules 1 and 2. The ARDSS will be implemented using Cool: Gen CASE tools and MS-SQL Server. The chapter will also discuss the management implications of the ARDSS. Finally, the relevant to the DSS development study objectives will be discussed within the chapter.

7-1 The ARDSS development

The proposed DSS methodology that was introduced in chapter five (Refer to section 5-13) will be adopted in the ARDSS development.

7-2 Discussion of the research objective No. 5 "*Use the proposed methodology to develop the required Admission and Registration DSS*"

This is an important objective because the proposed DSS methodology is going to be applied to the ARDSS development. This objective has been met by implementing the methodology's four modules:

1. **Module 0**. Needs' Analysis;
2. **Module 1**. Building the data warehouse;
3. **Module 2**. Knowledge from the KDD process;
4. **Module 3**. Building the DSS.

The details of implementing these four modules will be analyzed in the next sections.

7-3 Module 0: Needs' Analysis

This module has been accomplished in four phases as follows:

1. The first phase is the development and validation of a new research questionnaire that is used to define the current Admission and Registration information systems in the Egyptian Universities, and to explore the requirements that are not satisfied by these current systems;
2. The second phase the questionnaire was used to collect data from the Admission and Registration Managers in the Egyptian Universities;
3. The third phase, the Managers' information needs that are required to be satisfied (Refer to chapter six for more details) have been identified;
4. The last phase, Cool: Gen CASE tools Planning and Analysis phases were utilized to start the development. The development in Cool: Gen is based on the information needs, so in the next sections the information needs identification is discussed then Cool: Gen CASE tools planning and Analysis phases begin.

7-3-1 Information needs identification

The information needs identification is based on two factors:

1. The decisions that the Managers require the ARDSS to take;
2. The availability of sample data records to implement some of the decisions which are based on historic data. For example decision No. (**Q**) "Forecast course booking" is based on the availability of the following data: course title, course booking, course results, and number of students on each major per semester. However, none of the Egyptian Universities agreed to provide this data which means that the decision will not be implemented.

The following table (7-1) evaluates the decisions that the ARDSS will take based on the analysis of the questionnaire provided in chapter six (Refer to section 6-14-2). The Managers have identified *seventeen* decisions that the ARDSS can take. The decisions' variables required to take these decisions are included. Also the availability of sample data records is included in the last column.

No.	Decision	Decision variable(s)	Data available (Y/N)
A.	Accept or reject a new applicant	-High school percent -High school year -Age -Interview result -Gender -Abandoned before -Major requirement -Batch ceiling	Y
C.	Predict the new applicants that will join the faculty/ college/ institute this term/year based on our archival records	-Semester -Number of previous applicants	Y
D.	Predict the new applicants that will join the college this term/ year based on government statistics on secondary school students	-Semester -Government statistics -Market share	N
E.	Predict the new applicants that will join the college this term/ year based on our archival records besides other records like the government statistics	-Semester -Number of previous applicants -Government statistics -Market share	N
F.	Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major	-High school percent -High school certificate -High school origin -Age -Major -GPA -Gender -Nationality	Y
G.	Hold the applicant until the following term/year	-High school percent -High school year -Age -Interview result -Gender -Abandoned before -Major requirement -Batch ceiling	Y

H.	Accept or reject the applicant who is transferred from another educational institution	-High school percent -High school year -Age -Interview result -Gender -Abandoned before -Major requirement -Batch ceiling -Transferred from	Y
I.	Accept or reject the applicant who is transferred from another educational institution based on our transfer history records	-High school percent -High school year -Age -Interview result -Gender -Abandoned before -Major requirement -Batch ceiling -Transferred from -Transfer history	N
J.	Predict a student's performance based on the students' history we keep	-High school percent -Major -GPA	Y
K.	Predict a course's results based on the courses' history we keep	-Course title -Major -Exam results -Registration	N
L.	Classifying students into similar groups	-High school type -High school year -High school percent -DOB -Year in -Year out -GPA -Major -Nationality -Gender	Y
M.	Predict a student's performance based on the group that he/she belongs to	-High school type -High school year -High school percent -DOB -Year in -Year out -GPA -Major -Nationality -Gender	Y
N.	Set the student status to "On probation"	-GPA	Y

O.	Predict the "On probation" students based on the students' history we keep	-Current GPA -Past GPA -Major	N
P.	Make relationships between students' performance and academic departments	-Students' GPA -Major -Department	Y
Q.	Forecast course booking	-Course title -Course booking -Course results -Number of students on each major per semester	N
R.	Decide on Student abandonment	-GPA -On Probation status -Penalties	Y

Table (7-1). The decisions, their variables, and the availability of sample records.

Based on the information presented in table (7-1), only ***eleven*** decisions will be implemented in the ARDSS. *The following **seven** decisions will not be implemented*:

- Decision (**B**) was eliminated according to the users' requirements (i.e. users do not accept the ARDSS to take this decision);
- Decisions (**D, E, I, K, O, Q**) will not be implemented because their implementation requires historic data which none of the Universities agreed to provide this data.

The following table (7-2) includes the DECISIONS that the ARDSS will be built to take.

No.	Decision
A.	Accept or reject a new applicant
C.	Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records
F.	Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major
G.	Hold the applicant until the following term/year
H.	Accept or reject the applicant who is transferred from another educational institution
J.	Predict a student's performance based on the students' history we keep
L.	Classifying students into similar groups
M.	Predict a student's performance based on the group that he/she belongs to
N.	Set the student status to "On probation"
P.	Make relationships between students' performance and academic departments
R.	Decide on Student abandonment

Table (7-2). The Information Needs.

7-3-2 The use of Cool: Gen CASE tools "*How the DSS meets the users' information needs*"

The decisions identified in table (7-2) are the foundations of the INFORMATION NEEDS MATRIX that is to be created at the very early stage of the planning phase in Cool: Gen CASE tools. The INFORMATION NEEDS MATRIX drives the remaining CASE tools development components and phases including:

1. Defining the BUSINESS SYSTEM[1] that addresses a certain management problem to meet specific objectives. In this phase the Admission and Registration Decision Support System (ARDSS) was defined as the BUSINESS SYSTEM that needs to be constructed;
2. Identifying the FUNCTIONS of BUSINESS SYSTEM. The ARDSS addresses two main functions; Admission and Registration;

[1] For a definition of the Cool: Gen CASE tools components refer to the Glossary of terms.

3. Establishing relationships between the FUNCTIONS and the INFORMATION NEEDS. The functions have been detailed into sub functions; each of the sub function will meet one or more of the information needs;
4. Finding out the SUBJECT AREAS which are required to build the system FUNCTIONS. A number of subject areas have been identified to build each of the functions and sub functions. E.g. are course, students, scholarships, ...etc;
5. Create the ENTITY TYPES which are required to implement the SUBJECT AREAS. A number of entity types were defined and created to implement the subject areas. E.g. are batch, semester, nationality, tuition, major...etc;
6. Transfer the business logic into referential integrity (RI's) constraints and other constraints. The business logic has transferred to the system. E.g. what are the Admission requirements, when to grant scholarships, how the GPA is calculated...etc;
7. Defining the ELEMENTARY PROCESSES which are required to maintain the ENTITY TYPES. For each entity type a number of elementary processes have been defined. E.g. are add student, read student, update student, delete student...etc.
8. The DESIGN phase;
9. The GUI;
10. Finally, the generation and packaging.

Refer to Appendix (**F**) for more details.

7-4 Module 1: Building the data warehouse

In this module the University DW was designed and implemented on MS SQL Server.

7-4-1 Discussion of the research objective No. 2-2 "Designing the DW"

This research objective has been achieved by the undertaking the module in the following five steps:

1. *Study and evaluate the data sources;*
2. *Establish the source-to-target fields' matrix as a design validation tool;*
3. *Build and the DW Star Schema design using MS SQL Server;*
4. *The DW loading and updating strategies;*
5. *Design the Managers' reports using Crystal Reports.*

These five steps are analyzed in detail in the next section.

7-4-2 The University Data Warehouse design

1. *Study and evaluate the data sources*

In this step two procedures been carried out; firstly studying a DB[2] which is currently running in one of the Admission and Registration IS at an Egyptian University (i.e. OLTP). This DB was studied to determine the degree of support it is able to provide to meet the information needs of the users which have been identified by the questionnaire. Results of this procedure have shown that the DB is unable to respond to all the information needs and hence a decision was made to design a new DB. Secondly, the new DB has been designed which offers a complete support to the users' information needs. This newly designed DB will be the DSS DB based upon which the DW tables have been created. Table (7-3) shows the transformation process between the ARDSS DB entities and the DW tables.

[2] This DB's cannot be listed here in details because there is no permission for this.

ARDSS DB Entities	DW Tables	
1.COLLEGE 2.DEPARTMENT 3.MAJOR	1.COLLEGE	DIMENSION
4.APPLICANT 5.NATIONALITY 6.CERTIFICATE 7.UNIVERSITY	2.APPLICANT	DIMENSION
8.BATCH 9.SEMESTER	3.SEMESTER	DIMENSION
10. COURSE 11. PREREQUISITE 12. COURSE_MAJOR 13. LAB	4.COURSE	DIMENSION
14. ASSISTANTSHIP	5.ASSISTANTSHIP	DIMENSION
15. PENALTY	6.PENALTY	DIMENSION
16. STUDENTS 17. AUTHORITY 18. STUDENT_ASSISTANTSHIP 19. STUDENT_PENALTY 20. GPA	7.STUDENT	DIMENSION
21. TUITION	8.TUITION	DIMENSION
22. PAYMENT	9.PAYMENT	DIMENSION
23. REGISTRATION	10. REGISTRATION	DIMENSION
24. EXAM	11. EXAM	DIMENSION
25. MARK	12. MARK	DIMENSION
	13. STUDENT RECORD	FACT

Table (7-3). DW transformation process.

It is obvious that the TIME dimension is not part of the University DW. *A decision was made not to include the TIME dimension because of the following*:

a. The nature of the Admission and Registration functions. That is, the transactions do not occur on daily basis like a bank, stock exchange, or a super market. Instead the majority of the Admission and Registration transaction

happen by semester e.g. receiving applications, students' graduation, and course registration;

b. The reports required by the Admission and Registration are on semester basis, not on daily or even monthly basis;

c. The TIME dimension is a debatable point in literature. While some authors recommend that dimension to be part of every DW (Humphries, et al., 1999; Firestone, 1998; Hadden, 1998a), others (Kimball, 1996) said that if the business uses the dates and time spans on a year or month basis, then in this situation the TIME dimension could be deleted.

The Granularity[3] of the Fact table. Individual Student Record *BY* Applicant *BY* College *BY* Semester *BY* Course *BY* Assistantship *BY* Penalty *BY* Student *BY* Tuition *BY* Payment *BY* Registration *BY* Exam *BY* Mark.

2. *Establish the source-to-target fields' matrix as a design validation tool*

Humphries, et al. (1999) suggested the use of the source-to-target fields' matrix. The source-to-target fields' matrix is found in Appendix (D).

3. Build and the DW Star Schema design using MS SQL Server

The DW schema design is found in the following figure (7-1). The DW CREATE Statements are found in Appendix (D).

[3] According to Kimball (1996: 22) "Typical grains are individual transactions".

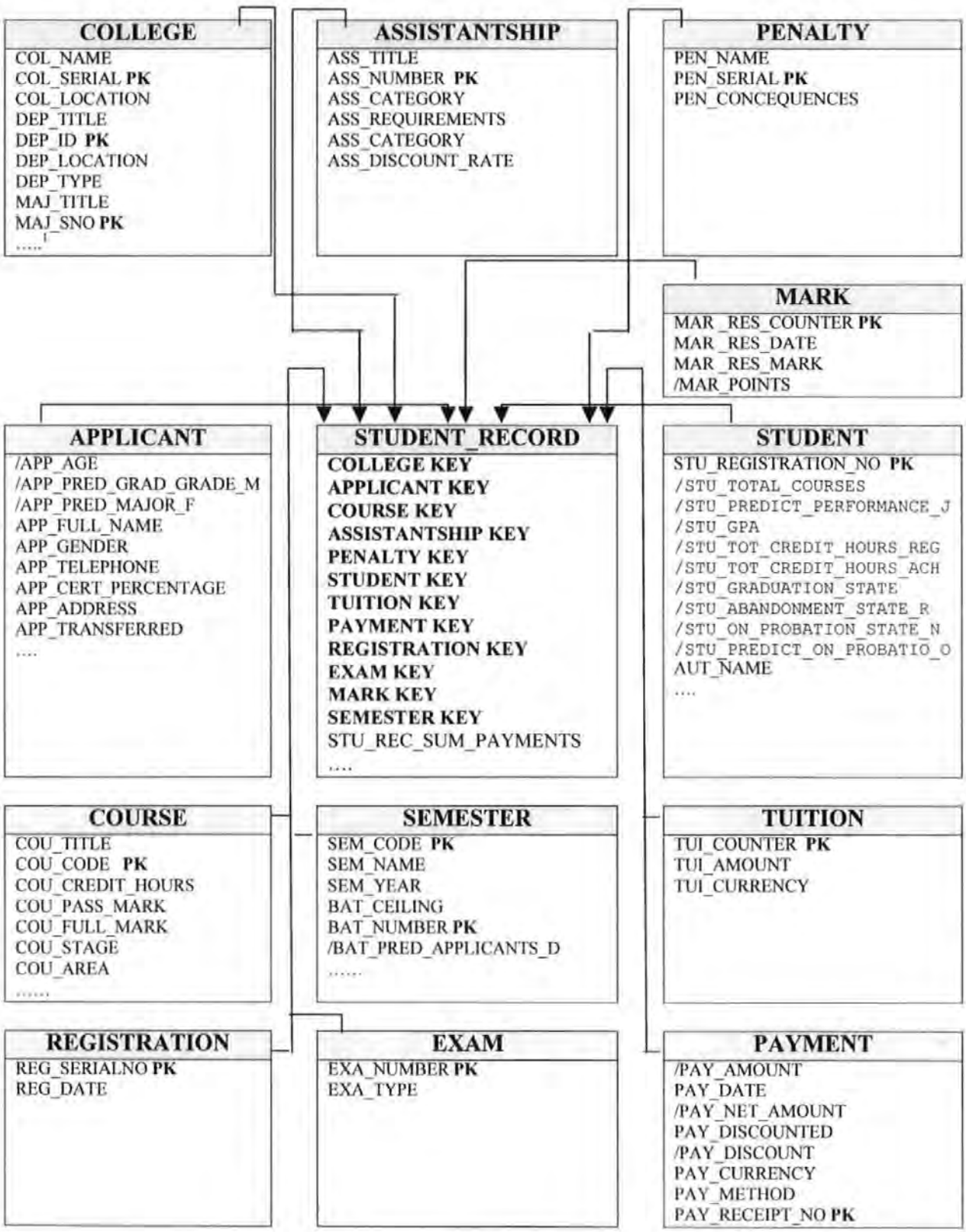


Figure (7-1). The University DW Star Schema.

[1] Refer to the source-to-target fields' matrix in Appendix (D) for the complete list of fields for each table.

4. The DW loading and updating strategies

The DW loading and updating strategies are described in the following sections.

A- Loading:

1. Two devices have been initiated on MS SQL Server; a DW data device and a DW dump device. Details are found in the following table (7-4):

Device name	Physical name	Description
ARDSS_DW	c:\ntragal\development\ardss_dw.dat	special, physical disk, 20 MB
ARDSS_DW_D	c:\dump\ardss_dw_d.dat	disk, dump device

Table (7-4). DW Creation details.

There are two other devices that have been initiated previously; the ARDSS DB and its dump device. The ARDSS DB (i.e. IEFDB) and dump devices are physically stored in different files for performance purposes (i.e. storing both DW and ARDSS DB in the same device would negatively affect their performance);

2. The DW dimension and fact tables (Refer to the Source-to-target fields' matrix in Appendix D) have been created in the ARDSS_DW (Refer to the CREATE statements in Appendix D);
3. The ARDSS_DW has been dumped to the ARDSS_DW_D (i.e. backup).

B-Updating:

1. *The incompatibility problem.* Release 5.0 of Cool: Gen that has been used for the development of the ARDSS can target MS SQL Server release 6 or 6.5 only. However, the two releases are not supported with any data warehouse features. Starting from MS SQL Server release 7.0 the Server has supported data warehousing and OLAP features (Compaq: Sizing Compaq Proliant Servers for Microsoft SQL Server 7.0 Data Marts, 1999; MCDBA SQL Server 7.0 Administration Study Guide, 1999; Sørensen and Alnor, 1999; McGehee, et al., 1998). According to Sørensen and Alnor paper in creating DW using SQL 7.0 (1999: 10-3) "there are several options when you want to load tables. The

most appealing way to load the schema is to build a data transformation service package (DTS Package) using the DTS Designer.";

The *Database/Object Transfer*[4] *and Bulk Copy Program (BCP)*[5] *tools* that are provided with MS SQL Server 6.5 (the release used for the development) are incapable of updating the data warehouse. The reason for this is that these tools can only move entire table(s) (i.e. all columns) and/or objects from one DB to another. However, this is not the case in data warehousing because not all source table columns are required to move and sometimes tables have been denormalized into one table.

2. *Resolving the incompatibility problem.* As release 7.0 of MS SQL Server is not used, and release 6.5 does not have the tools that could be used for updating the data warehouse, the data warehouse updating process have been undertaken using the following procedure:

    a. An ODBC connection that refers to the IEFDB (i.e. ODS) has been established on the client that points to the MS SQL Server;

    b. Data and Structure have been copied from the IEFDB to a new ODBC-compliant DB in MS ACCESS 2000;

    c. A CREATE queries have been used to denormalize the source tables into DIMENSION tables. For example the SEMESTER DIMENSION table has been created from the Semester and Batch source tables as follows:

```
SELECT dbo_SEMESTER.SEM_NAME, dbo_SEMESTER.SEM_YEAR,
dbo_SEMESTER.SEM_CODE, dbo_BATCH.FK_SEMESTERSEM_CODE,
dbo_BATCH.BAT_TITLE, dbo_BATCH.BAT_NUMBER, dbo_BATCH.BAT_CEILING,
dbo_BATCH.BAT_YEAR, dbo_BATCH.BAT_OPEN_DATE,
dbo_BATCH.BAT_CLOSING_DATE, dbo_BATCH.BAT_MARKET_SHARE,
dbo_BATCH.BAT_GOVERNMENT_STAT INTO SEMESTER_DIMENSION
FROM dbo_SEMESTER LEFT JOIN dbo_BATCH ON dbo_SEMESTER.SEM_CODE =
dbo_BATCH.FK_SEMESTERSEM_CODE;
```

    Also the COLLEGE DIMENSION table has been created from the College, Department, and Major source tables as follows:

[4] From one MS SQL Server to another.
[5] From one MS SQL Server to another or other applications.

```
SELECT dbo_COLLEGE.COL_NAME, dbo_COLLEGE.COL_SERIAL,
dbo_COLLEGE.COL_LOCATION, dbo_DEPARTMENT.DEP_TITLE,
dbo_DEPARTMENT.DEP_ID, dbo_DEPARTMENT.DEP_LOCATION,
dbo_DEPARTMENT.DEP_TYPE, dbo_DEPARTMENT.FK_COLLEGECOL_SERIAL,
dbo_MAJOR.MAJ_TITLE, dbo_MAJOR.MAJ_SNO,
dbo_MAJOR.MAJ_MIN_HIGH_SCHOOL_PERCENT,
dbo_MAJOR.FK_DEPARTMENTDEP_ID INTO [COLLEGE DIMENSION]
FROM dbo_COLLEGE LEFT JOIN (dbo_DEPARTMENT LEFT JOIN dbo_MAJOR ON
dbo_DEPARTMENT.DEP_ID = dbo_MAJOR.FK_DEPARTMENTDEP_ID) ON
dbo_COLLEGE.COL_SERIAL = dbo_DEPARTMENT.FK_COLLEGECOL_SERIAL;
```

d. After running the CREATE queries any columns could be deleted if not required by the data warehouse design;

e. A primary key column has been added to each DIMENSION table. The column type could be designated as Autonumber so the records generated can be numbered automatically (i.e. no data entry required), or any other data type;

f. The ODBC connection with the IEFDB that was created in step (a) is used whenever and update is to be performed (i.e. every semester or every year). When this happens the old tables are not overwritten or deleted;

g. The FACT table is created using intermediate tables, and then the relationships are established with the DIMENSION tables;

h. Loading data to the FACT table has been done using intermediate tables. CREATE QUERIES, SELECT, INNER JOIN, and JOIN statements have been used in this process. For example the COURSE and COLLEGE DIMENSION tables have been joined in an intermediate table, another intermediate table was created to join the data of both the APPLICANT and STUDENT DIMENSION tables..etc. Finally the intermediate tables are used to create the FACT table. The following statement is used for the creation of COURSE and COLLEGE intermediate table:

```
SELECT [COLLEGE DIMENSION].[college key], COURSE_DIMENSION.[course key],
[COLLEGE DIMENSION].COL_NAME, [COLLEGE DIMENSION].COL_SERIAL,
[COLLEGE DIMENSION].DEP_TITLE, [COLLEGE DIMENSION].DEP_ID, [COLLEGE
DIMENSION].MAJ_TITLE, [COLLEGE DIMENSION].MAJ_SNO,
COURSE_DIMENSION.COU_CODE, COURSE_DIMENSION.COU_TITLE,
```

COURSE_DIMENSION.LAB_TITLE, COURSE_DIMENSION.LAB_CODE, COURSE_DIMENSION.PRE_SNO, COURSE_DIMENSION.PRE_DETAIL1, COURSE_DIMENSION.COU_MAJ_COUNTER INTO INTERMEDIATE_COL_COU FROM [COLLEGE DIMENSION] INNER JOIN COURSE_DIMENSION ON [COLLEGE DIMENSION].MAJ_SNO = COURSE_DIMENSION.FK_MAJORMAJ_SNO;

i. Another ODBC connection has been established with the MS SQL Server to copy data and structure back from the MS ACCESS 2000 DB to the ARDSS_DW DB;

j. A TASK has been created on MS SQL Server. The task runs every semester, the role of the task is to delete the old copy of the data and table before the updated copy is to be done (or could be to move them to different DB for archiving purposes depends on the archiving strategy of the business). The TASK is to be scheduled on the same day of UPDATE, so that there is no time lag between the two. If a time lag has happened a query and/or report will generate an error because no tables exist. The TASK runs automatically without user intervention, after the task is performed it reports to the WIN NT EVENT VIEWER LOG file, and it is also set to e-mail the DW Administrator in cases of either success or failure;

k. Queries and reports could be run on the ARDSS_DW DB on the MS SQL Server, or alternatively on the MS ACCESS 2000 DB;

3. MS ACCESS 2000 restrictions. Two restrictions have been found in this suggested updating procedure; one is minor and the other is major.

a. The minor restriction happens when the DIMENSION table has been created and the data warehouse designer decides to use an Autonumber primary key (i.e. SURROGATE KEY). Then the intermediate queries will not be created because MS ACCESS can only display one Autonumber column in the output of a query. This problem has happened and a decision was made not to use a surrogate keys;

b. The major problem happens when the primary key of a certain table is composite (i.e. more than one column) in such a situation MS ACCESS can only handle ten column in a composite key. This problem has happened in this data warehouse. The FACT table primary key consists of twelve columns, only ten of which have been used as a composite primary key to the table to comply with the restriction.

5. *Design the Managers' reports using Crystal Reports*

Kimball (1996: 5) said, "Reporting is the primary activity in a data warehouse." Based upon Kimball's statement on reporting, meetings with some Managers in both Government and Private Egyptian Universities were conducted to identify the type of reports they want the DW to provide. A number of reports have been generated using SQL statements and presented in visualization techniques. Crystal Reports was used as a report generation tool. Copies of these reports are found in Appendix (D).

7-4-3 The discovered knowledge by the University DW reports

The University DW is able to generate the following reports (Details of the reports are found in Appendix D):

1. Nationality, Majors, and GPA;
2. Years in University, Majors, and GPA;
3. Majors versus Gender;
4. Age, GPA, and Majors;
5. High Schools, Majors, and GPA;
6. Majors value-added;
7. University value-added;

8. Demand Curve;
9. Gender Distribution;
10. Major Distribution;
11. Applicants' High School Scores, Majors, and Average Graduation Scores.

The reports were evaluated by the following:

1. Three managers in private Universities (Egypt);
2. Two managers in government Universities (Egypt);
3. One manager in the UK. Given that the DW reports are primarily created based on the requirements of the Egyptian managers, this step was undertaken because the Admission and Registration functions taking place in the two countries have some similarities, and the ARDSSQ was also pilot tested in the UK.

The Admission and Registration managers interviewed to evaluate the reports emphasized the following:

1. The reports are based on a long time span (i.e. 10 years) which enable them to look at the history to find any particular patterns for example:
    a. The *University and Departments' value added reports* could be useful for use by the managers to evaluate the performance, to compare between departments, to set the acceptance regulations (i.e. to increase or decrease minimum high school percent for the new applicants);
    b. The *demand curve report* could be used as an indicator for the marketing efforts exerted, could also be used to either increase or decrease the marketing budget, to compare the number of applicants with competitors, or could help identifying the market share of the University;

c. The *applicants' high school scores, majors, and average graduation scores* report is useful in establishing relationships between the various high school certificates in the different majors and the final graduation grades. This could enable the managers to evaluate the different certificates and to increase or decrease the number of students from certain certificates according to the past achievement;

2. They are unable to get these reports from their current Admission and Registration information systems. All of those who interviewed reported that neither one of the reports can be obtained from their current systems, moreover, when they ask for ad hoc reports they have to wait a long time to get it;
3. The reports enable them to reveal relationships which they are interested in and will help them making better decision;
4. It was also evident that different managers have different interpretations to the contents of the reports.

7-5 Module 2: Knowledge from the KDD process

In this module the KDD process has been applied to 1800 records. SQL, Visualization, and Clustering analysis techniques have been used as data mining techniques. The techniques have been applied for the following reasons:

a. Describing and representing the sample;

b. Finding the knowledge which will be stored in the ARDSS knowledge base;

c. Creating the managers' reports from the DW;

The following table (7-5) illustrates which data mining technique will be used to take which decision(s). The table is based on the literature reviewed in chapter four, the analysis of the techniques provided in chapter five and the decision variables presented in table (7-1).

The Data Mining technique	Goal (G)/ Task (T)	Use in related work	Use in the DSS
SQL	-(G): Description -(T): Summarization	-For data summarization -Finding shallow knowledge -Provide general statistics -Helping users to take structured decisions -Produce reports -Answers to FAQ	-Provide general statistics -Produce some reports -Data retrieval from the DW -Decisions: **A, C, G, H, J, N, P,** and **R.**
Visualization	-(G): Description -(T): Summarization	-For data summarization -Has the presentational advantage -To represent general statistics on data	- Represent general statistics -Finding shallow and multi-dimensional knowledge[6] -Used in some DW reports - To represent the output of the SQL
Clustering analysis	-(G): Description Prediction -(T): Clustering	-Used when data variables are inter-correlated	-Data description -Data prediction -Can handle large samples defined in many variables -Decisions: **F, L,** and **M.**

Table (7-5). The data mining techniques and the ARDSS decisions.

[6] Refer to Appendix (D).

7-5-1 The 2000 records sample description

Although **thirteen** Egyptian Universities (six private, seven government) participated in the needs analysis (i.e. **Module 0**), the knowledge base of the proposed ARDSS is based on sample data records drawn from the AASTMT students' database. The knowledge base is based on records from one University because of the following reasons:

1. Different Universities have different business logic. That is, the decision to accept or reject an applicant is taken by all the Universities, however, different conditions apply e.g. in the Faculty of Engineering at Alexandria University the minimum high school percentage accepted is 90%, whilst in the Faculty of Engineering at the AASTMT the minimum is 60%. Therefore, the needs of both Universities are the same (i.e. to take the accept/reject decision), whilst the business logic is different;
2. Also, different Universities have various attributes used to describe students' records;
3. The choice of the AASTMT is based on the following:
    a. They accepted to provide a reasonably large sample (i.e. 2000 records);
    b. They also allowed the researcher to interview the Admission and Registration managers as well as reviewing the Admission and Registration documentations;
    c. The AASTMT is the largest private University in Egypt in terms of Student population;
    d. The other Universities refused to provide sample data.

2000 records were drawn from the AASTMT students' DB. The 2000 records were then classified into two categories; the first includes 1800 records (90%) that will be used for the purpose of finding knowledge, the second includes 200 (10%) records that will be used to test the output of the knowledge.

The 2000 records include students who joined the AASTMT from 1985-1994 and graduated from 1990-1999. Each student record consists of the following attributes:

High school certificate code, high school certificate year, high school certificate origin, high school certificate percent, graduation major, graduation grade, graduation date, date of birth, nationality, gender, batch number, and registration number. The following tables describe the sample.

Gender	Number	Percent
Male	1803	90
Female	197	10
	2000	**100%**

Table (7-6). Gender' distribution in the sample.

Grade	Number	Percent
Poor	57	3
Pass	552	28
Good	539	27
Very Good	540	27
Excellent	105	5
Very Good- Honor	40	2
Excellent- Honor	167	8
	2000	**100%**

Table (7-7). Grades' distribution in the sample.

Major	Number	Percent
BBA English section	102	5
BBA Arabic section	298	15
Bachelor of Maritime Transport	79	4
BTech. Electronics	99	5
BTech. Marine Eng.	54	3
Bachelor of Hotels and Tourism	77	4
Bachelor of Maritime	118	6
B.Sc. Computers	185	9
B.Sc. Electronics	516	26
B.Sc. Marine Eng.	433	22
B.Sc. Mechanical Eng.	39	2
	2000	**100%**

Table (7-8). Majors' distribution in the sample.

High School	Number	Percent
Thanwya Amma- Math	867	43
Thanwya Amma- Science	827	41
Thanwya Azhar	138	7
Preparatory Diploma	4	.2
Thanwya Amma- Arts	69	4
Thanwya Amma- New	2	.1
Thanwya Amma New- Science	1	.05
Thanwya Amma Old- Science	79	4
Thanwya Amma Old- Arts	1	.05
Thanwya Amma Old- Science	2	.1
Thanwya Amma Old- Math	10	.5
	2000	**100%**

Table (7-9). High Schools' distribution in the sample.

Nationality	Number	Percent
Jordan	461	23.05
Sudan	130	6.5
Syria	34	1.7
Iraq	6	.3
Palestine	199	9.95
Qatar	12	.6
Lebanon	9	.45
Libya	34	1.7
Egypt	952	47.6
Yemen	73	3.65
Kuwait	17	.85
Ethiopia	3	.15
Namibia	5	.25
Saudi	32	1.6
Gabon	1	.05
Eritrea	11	.55
Oman	2	.1
Kenya	1	.05
Italy	1	.05
Emirates	7	.35
Pakistan	2	.1
Cyprus	1	.05
Terkistan	2	.1
Indonesia	1	.05
Morocco	1	.05
Tunisia	1	.05
USA	1	.05
Canada	1	.05
	2000	**100%**

Table (7-10). Nationalities' distribution in the sample.

7-5-2 The discovered knowledge

The entire KDD process (Refer to chapter four, section 4-4 for more details) was adopted to reach the discovered knowledge as follows:

1. *Developing an understanding of the application domain.* The Admission and Registration function in the Egyptian Universities represent the application domain. The objective of the KDD process is to find knowledge that will enable the decision makers to better understand their businesses and to enhance the quality of decisions their make (i.e. the twelve decisions were identified in table 7-1);
2. *Creating a target data set.* A target data set was created and stored in a DW DB;
3. *Data preprocessing.* The following activity was carried out to put the data obtained in a ready-to-use format:
    a. *Translation.* Some of the attributes obtained were in English (e.g. registration number, graduation grade, high school percent), whilst the remaining was in Arabic (e.g. graduation major, graduation grade, graduation date). A translation was undertaken to keep the sample records in the English language;
4. *Data reduction and projection.* The following activities were undertaken:
    a. *Data reduction.* The two attributes batch number and registration number were excluded from the analysis because the Admission and Registration managers interviewed ascertained that these attributes are used for identification purposes only and hence no decision is taken based on neither of which;
    b. *Data preprocessing.* Using the available attributes, the following two attributes have been calculated and added:
        i. Age was calculated using date of birth;
        ii. Value add was calculated using high school certificate percent and graduation grade;

c. *Data projection.* Each decision is related to some variables/attributes (Refer to table 7-1);

5. *Choosing the data mining goal and task.* The goals of the data mining techniques have been identified as description and prediction, and the tasks are summarization and clustering;
6. *Choosing the data mining technique(s).* The chosen techniques are SQL, Visualization, and Clustering;
7. *Data mining.* Searching for patterns by applying the techniques to the 1800 records data set;
8. *Interpreting the information gained by the mining techniques.* The results of applying the techniques are interpreted to formulate relevant business logic. In practice, iterations from steps 1 to 7 often occur as new data is obtained from the data source(s), or if applying new data mining techniques;
9. *Consolidating the discovered knowledge.* Reporting the knowledge to the interested parties and checking the discovered knowledge with the previously known knowledge. The knowledge discovered by applying the techniques is described in the following rules:

The discovered knowledge is will be described in the form of rules in the following sections:

Rule No. 1

How obtained: This rule was obtained by two sources; interviewing[7] some Admission and Registration managers at the AASTMT and documentation review and partially validated by SQL statement running on the sample data set. The SQL validation was partially because the sample data set does not have all the attributes required to validate the rule e.g. interview.

Decision No. A

"Accept or reject a new applicant"

[7] Unstructured interviews took place with the AASTMT Admission Officer, Registrars, Deans, and other Senior Academics.

IF major = "Hotels and Tourism" AND interview= "Satisfactory" AND high school percentage >= 60 AND age < 25 AND number of applicants <= batch ceiling
THEN accept applicant = "Y"
ELSE IF major = "Maritime" OR "Marine Eng." AND gender= "Male" AND high school percentage >= 60 AND age < 25 AND number of applicants <= batch ceiling
THEN accept applicant = "Y"
ELSE IF major <> "Maritime" OR "Marine Eng." OR "Hotels and Tourism" AND high school percentage >= 60 AND age < 25 AND number of applicants <= batch ceiling
THEN accept applicant = "Y"
ELSE
accept applicant = "N"
END IF

Rule No. 2

How obtained: This rule was obtained by two sources; interviewing some Admission and Registration managers at the AASTMT and documentation review.

Decision No. C

"Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records"

IF semester= "September" THEN
applicants predicted = 2000
ELSE IF semester = "February" THEN
applicants predicted = 500
ELSE
applicants predicted = 0
END IF

Rule No. 3

How obtained: This rule was obtained by applying the Ward's clustering technique based on a Euclidean metric measure running on the sample data set.

Decision No. F[8]

[8]Refer to the CLUSTAN output in Appendix (F) for the Dendrogram.

"Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major"

Cluster 1: (38 members)

IF high school certificate="Thanwya Azhar" AND high school origin="Egypt" AND high school percent="66" AND nationality="Libya'' AND gender="Male"

THEN major="BBA Arabic section" AND age on graduation="23" AND GPA="Very Good"

Cluster 2: (81 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="65" AND nationality="Egypt'' AND gender="Male"

THEN major=" B.Sc. Electronics" AND age on graduation="25" AND GPA="Good"

Cluster 3: (136 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="65" AND nationality=" Oman'' AND gender="Male"

THEN major=" Bachelor of Hotels and Tourism" AND age on graduation="27" AND GPA="Very Good"

Cluster 4: (50 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="56" AND nationality="Egypt'' AND gender="Female"

THEN major=" BBA Arabic section" AND age on graduation="24" AND GPA="Good"

Cluster 5: (146 members)

IF high school certificate="Thanwya Azhar" AND high school origin="Egypt" AND high school percent="52" AND nationality="Egypt'' AND gender="Male"

THEN major=" Bachelor of Maritime Transport" AND age on graduation="25" AND GPA="Pass"

Cluster 6: (60 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="52" AND nationality="Egypt'' AND gender="Male"

THEN major=" BTech. Electronics" AND age on graduation="25" AND GPA="Poor"

Cluster 7: (42 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="55" AND nationality="Sudan'' AND gender="Male"

THEN major=" Bachelor of Maritime Transport" AND age on graduation="25" AND GPA="Good"

Cluster 8: (46 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="50" AND nationality="Syria'' AND gender="Male"
THEN major=" B.Sc. Electronics" AND age on graduation="26" AND GPA="Good"

Cluster 9: (32 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="59" AND nationality="Jordan'' AND gender="Male"
THEN major=" B.Sc. Marine Eng." AND age on graduation="26" AND GPA="Good"

Cluster 10: (66 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="58" AND nationality="Egypt'' AND gender="Male"
THEN major=" B.Sc. Electronics" AND age on graduation="25" AND GPA="Very Good"

Cluster 11: (206 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="51" AND nationality="Egypt'' AND gender="Male"
THEN major=" B.Sc. Computers" AND age on graduation="25" AND GPA="Good"

Cluster 12: (89 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="78" AND nationality="Lebanon'' AND gender="Male"
THEN major=" Bachelor of Maritime" AND age on graduation="24" AND GPA="Very Good"

Cluster 13: (78 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="73" AND nationality="Egypt'' AND gender="Male"
THEN major=" B.Sc. Electronics" AND age on graduation="24" AND GPA="Very Good"

Cluster 14: (78 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="66" AND nationality="Jordan'' AND gender="Male"
THEN major=" Bachelor of Maritime Transport" AND age on graduation="25" AND GPA="Good"

Cluster 15: (37 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Yemen" AND high school percent="68" AND nationality="Sudan'' AND gender="Male"
THEN major=" B.Sc. Electronics" AND age on graduation="25" AND GPA="Good"

Cluster 16: (128 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Libya" AND high school percent="74" AND nationality="Jordan'' AND gender="Male"

THEN major=" B.Sc. Computers" AND age on graduation="25" AND GPA="Very Good"

Cluster 17: (92 members)

IF high school certificate="Thanwya Amma- Science" AND high school origin="Egypt" AND high school percent="80" AND nationality="Jordan'' AND gender="Male"

THEN major=" B.Sc. Electronics" AND age on graduation="24" AND GPA="Very Good"

Cluster 18: (58 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="87" AND nationality="Libya'' AND gender="Male"

THEN major=" B.Sc. Computers" AND age on graduation="24" AND GPA="Very Good"

Cluster 19: (107 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="85" AND nationality="Saudi'' AND gender="Male"

THEN major=" Bachelor of Maritime" AND age on graduation="25" AND GPA="Excellent"

Cluster 20: (14 members)

IF high school certificate="Thanwya Amma- Math" AND high school origin="Egypt" AND high school percent="86" AND nationality="Jordan'' AND gender="Male"

THEN major=" B.Sc. Computers" AND age on graduation="24" AND GPA="Very Good"

Cluster 21: (133 members)

IF high school certificate="IGCSE- Old" AND high school origin="Eritrea" AND high school percent="76" AND nationality="Egypt"' AND gender="Male"

THEN major=" Bachelor of Hotels and Tourism" AND age on graduation="25" AND GPA="Good"

Cluster 22: (9 members)

IF high school certificate="Thanwya Amma Old- Science" AND high school origin=" Lebanon" AND high school percent="80" AND nationality='' Lebanon" AND gender="Male"

THEN major=" B.Sc. Computers" AND age on graduation="25" AND GPA="Very Good"

Cluster 23: (27 members)

IF high school certificate="Thanwya Amma Old- Science" AND high school origin="Libya" AND high school percent="63" AND nationality="Lebanon" AND gender="Male"

THEN major=" B.Sc. Computers" AND age on graduation="23" AND GPA="Excellent"

Cluster 24: (37 members)

IF high school certificate="Thanwya Amma Old- Math" AND high school origin="Egypt" AND high school percent="69" AND nationality="Libya'' AND gender="Male"

THEN major=" B.Sc. Computers" AND age on graduation="25" AND GPA="Good"

Rule No. 4

How obtained: This rule was obtained by two sources; interviewing some Admission and Registration managers at the AASTMT and documentation review and partially validated by SQL statement running on the sample data set.

Decision No. G

"Hold the applicant until the following term/year"

IF high school percentage >= 60 AND age < 25 AND number of applicants > batch ceiling

THEN deferred applicant = "Y"

ELSE

deferred applicant = "N"

END IF

Rule No. 5

How obtained: This rule was obtained by two sources; interviewing some Admission and Registration managers at the AASTMT and documentation review and partially validated by SQL statement running on the sample data set.

Decision No. H

"Accept or reject the applicant who is transferred from another educational institution"

IF major = "Hotels and Tourism" AND interview= "Satisfactory" AND high school percentage >= 60 AND age < 25 AND number of applicants <= batch ceiling AND previous abandonment = "N"

THEN accept applicant = "Y"

ELSE IF major = "Maritime" OR "Marine Eng." AND gender= "Male" AND high school percentage >= 60 AND age < 25 AND number of applicants <= batch ceiling AND previous abandonment = "N"

THEN accept applicant = "Y"

ELSE IF major <> "Maritime" OR "Marine Eng." OR "Hotels and Tourism" AND high school percentage >= 60 AND age < 25 AND number of applicants <= batch ceiling AND previous abandonment = "N"

THEN accept applicant = "Y"

ELSE

accept applicant = "N"

END IF

Rule No. 6

How obtained: This rule was obtained and validated by SQL statement running on the sample data set.

Decision No. J

"Predict a student's performance based on the students' history we keep"

IF high school percentage <= 65 THEN

predicted student performance = "Between 0-2"

ELSE IF high school percentage > 65 THEN

predicted student performance = "Between 3-6"

ELSE

predicted student performance = "Undefined"

END IF

Rule No. 7

How obtained: This rule was obtained by applying the Ward's clustering technique based on a Euclidean metric measure running on the sample data set.

Decision No. L[9]

"Classifying students into similar groups"

Cluster 1: (44 members)

high school type= "Thanwya Amma- Science" AND high school year= "90" AND year in= "91" AND high school percent= "73" AND major= "Bachelor of Maritime Transport" AND graduation date="96" AND GPA= "Very Good" AND DOB= "72"AND nationality= "Libya" AND gender= "Male"

Cluster 2: (141 members)

[9]Refer to the CLUSTAN output in Appendix (F) for the Dendrogram.

high school type= "Thanwya Amma- Math" AND high school year= "91" AND year in= "91" AND high school percent= "75" AND major= "B.Sc. Electronics" AND graduation date="97" AND GPA= "Very Good" AND DOB= "73"AND nationality= "Egypt" AND gender= "Male"

Cluster 3: (128 members)

high school type= "Thanwya Amma- Math" AND high school year= "88" AND year in= "88" AND high school percent= "80" AND major= "B.Sc. Computers" AND graduation date="94" AND GPA= "Very Good" AND DOB= "70"AND nationality= "Jordan" AND gender= "Male"

Cluster 4: (65 members)

high school type= "Thanwya Amma- Science" AND high school year= "88" AND year in= "88" AND high school percent= "73" AND major= "B.Sc. Marine Eng." AND graduation date="95" AND GPA= "Very Good" AND DOB= "70"AND nationality= "Jordan" AND gender= "Male"

Cluster 5: (116 members)

high school type= "Thanwya Amma- Math" AND high school year= "90" AND year in= "90" AND high school percent= "88" AND major= "B.Sc. Computers" AND graduation date="96" AND GPA= "Very Good" AND DOB= "72"AND nationality= "Lebanon" AND gender= "Male"

Cluster 6: (84 members)

high school type= "Thanwya Amma- Science" AND high school year= "88" AND year in= "89" AND high school percent= "88" AND major= "B.Sc. Electronics" AND graduation date="94" AND GPA= "Excellent" AND DOB= "70"AND nationality= "Jordan" AND gender= "Male"

Cluster 7: (79 members)

high school type= "Thanwya Amma- Science" AND high school year= "91" AND year in= "91" AND high school percent= "64" AND major= "BBA Arabic section" AND graduation date="96" AND GPA= "Very Good" AND DOB= "75"AND nationality= "Egypt" AND gender= "Female"

Cluster 8: (148 members)

high school type= "Thanwya Amma- Math" AND high school year= "91" AND year in= "91" AND high school percent= "63" AND major= "B.Sc. Computers" AND graduation date="97" AND GPA= "Good" AND DOB= "73"AND nationality= "Egypt" AND gender= "Male"

Cluster 9: (103 members)

high school type= "Thanwya Amma- Science" AND high school year= "92" AND year in= "92" AND high school percent= "57" AND major= "B.Sc. Computers" AND graduation date="97" AND GPA= "Good" AND DOB= "74"AND nationality= "Egypt" AND gender= "Male"

Cluster 10: (147 members)

high school type= "Thanwya Amma- Science" AND high school year= "91" AND year in= "91" AND high school percent= "56" AND major= "BBA Arabic section" AND graduation date="96" AND GPA= "Very Good" AND DOB= "73"AND nationality= "Egypt" AND gender= "Male"

Cluster 11: (115 members)

high school type= "Thanwya Azhar" AND high school year= "91" AND year in= "91" AND high school percent= "51" AND major= "BTech. Electronics" AND graduation date="97" AND GPA= "Good" AND DOB= "73"AND nationality= "Egypt" AND gender= "Male"

Cluster 12: (92 members)

high school type= "Thanwya Amma- Science" AND high school year= "90" AND year in= "90" AND high school percent= "55" AND major= "Bachelor of Hotels and Tourism" AND graduation date="97" AND GPA= "Good" AND DOB= "72"AND nationality= "Syria" AND gender= "Male"

Cluster 13: (45 members)

high school type= "Thanwya Amma- Science" AND high school year= "87" AND year in= "88" AND high school percent= "55" AND major= "Bachelor of Maritime Transport" AND graduation date="94" AND GPA= "Poor" AND DOB= "69"AND nationality= "Libya" AND gender= "Male"

Cluster 14: (60 members)

high school type= "Thanwya Amma- Math" AND high school year= "88" AND year in= "88" AND high school percent= "53" AND major= "B.Sc. Electronics" AND graduation date="95" AND GPA= "Good" AND DOB= "70"AND nationality= "Egypt" AND gender= "Male"

Cluster 15: (43 members)

high school type= "Thanwya Amma- Math" AND high school year= "89" AND year in= "90" AND high school percent= "74" AND major= "Bachelor of Hotels and Tourism" AND graduation date="96" AND GPA= "Good" AND DOB= "70"AND nationality= "Oman" AND gender= "Male"

Cluster 16: (21 members)

high school type= "Thanwya Amma- Math" AND high school year= "88" AND year in= "89" AND high school percent= "57" AND major= "Bachelor of Maritime" AND graduation date="95" AND GPA= "Good" AND DOB= "66"AND nationality= "Eritrea" AND gender= "Male"

Cluster 17: (75 members)

high school type= "Thanwya Amma- Science" AND high school year= "90" AND year in= "90" AND high school percent= "68" AND major= "BTech. Marine Eng." AND graduation date="96" AND GPA= "Good" AND DOB= "72"AND nationality= "Sudan" AND gender= "Male"

Cluster 18: (42 members)

high school type= "Thanwya Amma- Science" AND high school year= "87" AND year in= "87" AND high school percent= "58" AND major= "B.Sc. Marine Eng." AND graduation date="93" AND GPA= "Good" AND DOB= "68"AND nationality= "Jordan" AND gender= "Male"

Cluster 19: (74 members)

high school type= "Thanwya Amma- Science" AND high school year= "87" AND year in= "87" AND high school percent= "66" AND major= "B.Sc. Marine Eng." AND graduation date="94" AND GPA= "Good" AND DOB= "68"AND nationality= "Jordan" AND gender= "Male"

Cluster 20: (95 members)

high school type= "Thanwya Amma- Science" AND high school year= "87" AND year in= "87" AND high school percent= "65" AND major= "B.Sc. Electronics" AND graduation date="94" AND GPA= "Very Good" AND DOB= "68"AND nationality= "Egypt" AND gender= "Male"

Cluster 21: (9 members)

high school type= "Thanwya Amma New- Science" AND high school year= "90" AND year in= "90" AND high school percent= "76" AND major= "Bachelor of Hotels and Tourism" AND graduation date="97" AND GPA= "Good" AND DOB= "72"AND nationality= "Yemen" AND gender= "Male"

Cluster 22: (30 members)

high school type= "Thanwya Amma Old- Science" AND high school year= "91" AND year in= "91" AND high school percent= "79" AND major= "B.Sc. Electronics" AND graduation

date="97" AND GPA= "Very Good" AND DOB= "74"AND nationality= "Palestine" AND gender= "Female"

Cluster 23: (34 members)

high school type= "Thanwya Amma Old- Science" AND high school year= "91" AND year in= "91" AND high school percent= "62" AND major= "B.Sc. Computers" AND graduation date="97" AND GPA= "Good" AND DOB= "73"AND nationality= "Libya" AND gender= "Male"

Cluster 24: (10 members)

high school type= "Thanwya Amma Old- Math" AND high school year= "91" AND year in= "91" AND high school percent= "69" AND major= "B.Sc. Computers" AND graduation date="97" AND GPA= "Good" AND DOB= "73"AND nationality= "Libya" AND gender= "Male"

Rule No. 8

How obtained: This rule was obtained by applying the Ward's clustering technique based on a Euclidean metric measure running on the sample data set.

Decision No. M

"Predict a student's performance based on the group that he/she belongs to"

Depends on Decision No. **L**.

For example: cluster 1 will be represented as:

IF high school type= "Thanwya Amma- Science" AND high school year= "90" AND year in= "91" AND high school percent= "73" AND major= "Bachelor of Maritime Transport" AND graduation date="96" AND DOB= "72"AND nationality= "Libya" AND gender= "Male" THEN GPA= "Very Good"

Rule No. 9

How obtained: This rule was obtained by two sources; interviewing some Admission and Registration managers at the AASTMT and documentation review and validated by SQL statement running on the sample data set.

Decision No. N

"Set the student status to 'On probation'"

IF student GPA < 2.0 THEN

student status = "On probation"

ELSE

student status = "Normal"

END IF

Rule No. 10

How obtained: This rule was obtained and validated by SQL statement running on the sample data set.

Decision No. P

"Make relationships between students' performance and academic departments"

m = major

students performance = AVERAGE graduation grade for m

WHERE 12 > m > 0

IF 1> students performance >= 0 THEN

major performance = "Poor"

ELSE IF 2> students performance >= 1 THEN

major performance = "Satisfactory"

ELSE IF 3> students performance >= 2 THEN

major performance = "Good"

ELSE IF 4> students performance >= 3 THEN

major performance = "V. Good"

ELSE IF 5> students performance >= 4 THEN

major performance = "Excellent"

ELSE IF 6> students performance >= 5 THEN

major performance = "V. Good, Honor"

ELSE IF students performance = 6 THEN

major performance = "Excellent, Honor"

ELSE

major performance = "Undefined"

END IF

END WHILE

Rule No. 11

How obtained: This rule was obtained by two sources; interviewing some Admission and Registration managers at the AASTMT and documentation review and partially validated by SQL statement running on the sample data set.

Decision No. R

"Decide on Student abandonment"

IF penalty = "Y" AND student GPA < 2.0 THEN

student abandoned = "Y"

ELSE IF penalty = "Y" AND student GPA >= 2.0 THEN

student abandoned = "W"

ELSE

student abandoned = "N"

END IF

The discovered knowledge is considered *deep knowledge* because of the following:

1. More than one mining techniques was used:
    a. SQL which is able to reach the shallow knowledge;
    b. Visualization which can explore the multi-dimensional knowledge;
    c. Clustering which enables the hidden knowledge to be identified;
2. The entire KDD process was adopted (i.e. from domain understanding till the knowledge consolidation), instead of just applying the techniques;
3. The knowledge rules evaluated by the Admission and Registration managers were described as deep and new. Same Admission and Registration managers who were involved in the evaluation of the University DW reports evaluated the discovered knowledge, they reported positively on the validity of the knowledge;
4. The use of the DW added a strategic dimension to the discovered knowledge.

7-6 Module 3: Building the ARDSS

According to the classifications of DSS that were discussed in chapter two (Refer to chapter two; section 2-6-6 for more details), the ARDSS is a *rule-oriented* DSS. The reason for that classification is that the decisions that the ARDSS takes are based on rules in the form of IF..THEN statements, which is classified according to Turban and Aronson, 1998; Holsapple and Whinston, 1996 as rule-oriented DSS. In the following sections the ARDSS components, testing, installation, and implications are elaborated.

7-6-1 The ARDSS components

Basically any DSS could have up to five components including: data management subsystems, model management subsystems, knowledge management subsystems, user interface subsystems, and the systems users (Refer to chapter two, section 2-6-4 for more details). However, Marakas, 1998; Holsapple and Whinston, 1996 reported that not all of these components are found on every DSS. For instance, Holsapple and Whinston, 1996 mentioned that the existence of the knowledge management subsystems is optional. Furthermore, Chen and Sinha (1996) in their inventory DSS they adopted an object-oriented approach based upon which they identified three components: data management, model management and user interface. Also, Raghunathan (1996) identified two DSS components only; model management and data management. Raghunathan did not include a user interface component, however, he emphasized the importance and effect of using CASE tools in DSS development (1996: 309) "We believe that a design tool such as CASE for DSS will significantly enhance the use of DSS".

For the purpose of the ARDSS, the following components have been identified:

1. *Data management component.* In this component the ARDSS DB has been created in Cool: Gen CASE tools and being transferred to MS-SQL Server (Refer to the DB CREATE statements in Appendix (F));
2. *Knowledge management component.* The results of applying the knowledge discovery techniques were stored using the design tool provided by Cool: Gen CASE tools. As new data is added to the DW, the knowledge discovery techniques run again and new knowledge could be found, which could affect the knowledge base by adding new rules, changing or modifying or deleting existing rules. The ARDSS knowledge base is detailed in section 7-5-2;
3. *User Interface.* The ARDSS adopts a GUI environment, a GUI example is illustrated in Appendix (F);
4. *Users.* The users of the ARDSS have been identified in chapter six (Refer to section 6-3) as (Dean, Associate Deans, Registrars, Admission Officers, and Others).

It is worth mentioning that the ARDSS was designed in a client/server environment. The server running Windows NT (4.0), and MS-SQL Server (6.5). On the Client side, Cool: Gen CASE tools (5.0), CLUSTAN Graphics (5.0), and Crystal reports (4.6). The server is an Intel P 266 MHZ, 64 MB RAM, and the Client was an Intel P 233 MHZ, 64 MB RAM, and Windows (98) OS.

7-6-2 Testing the ARDSS

Software testing is a fundamental component in all software development approaches. It is evident that many software testing methods and techniques are available. However, there is no single ideal software testing technique and/or method for assessing software quality, rather a combination of testing techniques and/or methods would be employed (Parrish and Zweben,

1991; Weyuker, 1986). Crispin (2001) said that the goal of testing is not to cover 100% of every single path of the code, instead is to do the minimum testing which ensures the required business value has been delivered to the customers.

The ARDSS has passed four levels of testing:

1. *The consistency check level provided by Cool: Gen CASE tools.* No errors or warnings received. The following message received from Cool: Gen after running the consistency check at all the development levels (i.e. Planning, Analysis, Design, and Construction);

```
ardss: Window Code Generation

Lines   Status    Type           Name
32915 Completed   Triggers       REFERENTIAL INTEGRITY TRIGGERS
32807 Completed   Window Manager GUIMENU1
418   Completed   Install Deck   REFERENTIAL INTEGRITY TRIGGERS
2667  Completed   Install Deck   GUIMENU1

68807 lines of code generated at 67679 lines/minute.
  All modules generated successfully.
```

2. *Professionals level.* The ARDSS has been sent to be checked at:
    a. *Computer Associates (CA) in London* (This company is the product owner and provides Cool: Gen technical support for the UK). Technical Support Department has provided some minor comments on the system. The comments have been taken into consideration. Then resent again with no corrections received.
    b. *Systems Integrators (SI) in Cairo* (This company provides Cool: Gen technical support for the Middle East Countries). No negative comments have been received.
3. *Users level.* Some Registrars and Admission Officers in Egyptian Universities have positively received the ARDSS functions and capabilities;
4. *Rules level.* The results of testing the ARDSS rules are analyzed in the following table (7-11). The sample used to test those decisions consists of 200 records. The 200 records used

to test the decisions were not part of the records used to build the knowledge base of the ARDSS. Moreover, all the majors are represented in the 200 records.

Decision number	Testing result
A	**64%**
C	-
F	**69%**
G	**100%**
H	**64%**
J	**60%**
L	**100%**
M	**60%**
N	**100%**
P	-
R	**100%**

Table (7-11). Results of testing the ARDSS rules[10].

The ARDSS testing results presented in table (7-11) reveal the following findings:

1. The testing results of the ARDSS decisions' rules are satisfactory. **Four** decisions (**G**, **L**, **N**, and **R**) have received **100%** testing scores. **Five** decisions received less than 100% including decisions (**A**; **64%**), (**F**; **69%**), (**H**; **64%**), (**J**; **60%**), and (**M**; **60%**);
2. The five decisions which received less than 100% testing results are due to the following:
    a. Three decisions of the five are used for prediction (i.e. **F**, **J**, and **M**);
    b. The effect of random sampling. For example major 9 has the highest prediction ability because it has the highest number of students on the sample (454 out of 1800 = 25.2 %). Conversely, majors 3, 5, and 11 have the lowest prediction rates because of their weak representation in the sample (71, 48, 46) respectively;

[10] Sometimes the rule to be tested requires fields that are not part of the sample data set (e.g. rule No.1 attribute *interview* is not part of the sample data set, also rules No.2, 5, 9, 11); in such situations artificial values were used.

c. Two decisions handle large number of variables. Decision **F** is based on *8 variables,* whilst Decision **M** is based on 10 *variables*;

d. For decisions (**A** and **H**), the testing scores are affected by the change of the Accept/Reject regulations. That is, the sample describes the students who joined the AASTMT from 85-95 during these years the minimum accepted high school percentage was 40%. However, the rules of decisions A and H reflect the current Accept/Reject regulations for which the minimum is 60%. In other words, if the minimum of rules no.1 and 5 became 40%, the testing results would have been 100% for both;

3. For decision **C**, the available data set is insufficient to be used for testing. The rule of decision C requires the number of students joined the University every term for a number of years to be known i.e. in September 90 the number is X, in February the number of students is Y, and it goes on for year 91, 92 etc. Since these data are not available, testing the decision rule was not carried out;

1. For decision **P**, testing the decision rule was undertaken in a different way because the rule is represented by averages. The results are satisfactory as the differences between averages are reasonable except for major no 11. Table (7-12) illustrates the results:

Major codes		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Average Grades	**1800 records**	3.04	2.15	6	1.86	2.1	2.63	3.32	2.54	2.45	2.74	2.67
	200 records	3	2.2	6	2.1	2.3	2.56	3.54	2.89	2.21	2.7	2
Absolute average differences		0.04	0.05	-	0.24	0.2	0.07	0.22	0.35	0.24	0.04	0.67

Table (7-12). Testing decision no.P; rule no.10.

7-6-3 The ARDSS installation

The ARDSS has been successfully built, generated, and tested in a local machine. However, the ARDSS has the features and capabilities to work in a client/server environment. The reasons for not installing the ARDSS in a real client/server environment are:

1. The release that was used for the development is an Academic release (Not Professional). This means that it does not have all the capabilities required for that kind of development. For example running the GUI in a client/server environment produced the following error:

> OPSYS Not Purchased.
> TPMON Not Purchased.
> Invalid parameters for generation.

2. The Academic release, despite missing some professional components, is not supposed to be used for generating commercial products (there is an 80 % discount on the price of the Academic release);
3. The researcher has contacted Computer Associates in London for this regard. Their officials said that installing and testing the ARDSS in a real environment would cost the researcher more than 1000 GBP per working day. It was assumed that this process would last for 3-5 days;
4. No response was received from SI in Cairo for this regard. The reason they did not reply is because the AASTMT (AASTMT is the researcher's employer and buyer of the Cool: Gen license) has ended the maintenance contract with them. As a result of that the researcher was unable to get such a service;
5. The researcher intends to upgrade the Academic release to a Professional one and also upgrade the MS-SQL release associated whenever funds available in future. When this happens the installation will take place;

6. The researcher also contacted Microsoft (in both UK and USA) for a free MS-SQL Server new release (i.e. Release 7.0), but unfortunately they denied the request.

7-6-4 The ARDSS limitations

1. The ARDSS is restricted to the knowledge stored in its knowledge-base;
2. The ARDSS is able to take only eleven decisions, which is turn means that not all of the Admission and Registration related decisions are incorporated into the system;
3. The ARDSS is an environment-specific system; that is it requires a Client/Server environment which has MS SQL Server 6.5 RDBMS running on a Windows NT 4 OS, Crystal reports 4.6, CLUSTAN graphics 5.0, and a Windows 95 or 98 on a Pentium machine;
4. The ARDSS is designed for the Egyptian Universities to be used by Deans, Associate Deans, Registrars, Admission Officers, and Others;
5. The discovered knowledge (Refer to section 7-5-2) is based on records drawn from the AASTMT students' DB. Although other Universities' managers including both Government and Private found the majority of the knowledge base relevant and acceptable, this knowledge can only be used for decision making at the AASTMT, and if any other University will use the ARDSS records from this University need to be included in the KDD process;
6. The ARDSS has not been installed (Refer to section 7-6-3).

7-7 The management implications of the ARDSS

1. *Gaining competitive advantages.* Robson, 1997; O'Brien, 1996; Keyes, 1993; Geiger, 1992 reported that organisations are able to achieve competitive advantages by utilizing information systems (Refer to chapter two, section 2-4 for more details). With regard to

the ARDSS, the following competitive strategies could be targeted and hence Universities are able to achieve competitive advantages:

a. Lower cost. Using the ARDSS is expected to reduce the transaction cost. Following are some examples:
    i. Accept/Reject applicant;
    ii. Decide on student abandonment;
    iii. Predicting current students' performance;
    iv. Generating reports that are based on archival data.
b. Differentiation. The Universities are able to get new features by using the ARDSS which are not currently available for the majority of their rivals. Following are some examples:
    i. Predicting the new and existing students' performance;
    ii. Predict on-probation state;
    iii. Storing historical data in a ready to use format for managers (i.e. DW).

2. *Managers are more committed and informed.* By using the ARDSS, Admission and Registration Managers are more committed to the system because their information requirements drive the entire development activities. Besides, the use of the KDD process and the DW technology make the managers better informed which enable them to understand their business better and to take high quality decisions.
3. *Better-served customers (i.e. students).* Geiger (1992: 6) said, "Universities and other non-profit organisations need to respond to an increasingly competitive climate and adopt enhanced financial management attitudes in an effort to win student clients"; by using the ARDSS students are able to get some useful information and hence increase their satisfaction. Following are some examples:

a. The ARDSS is able to make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major;

b. The ARDSS is also able to predict the student's GPA.

4. *The benefits of using CASE tools.* Baik, 2000; Cool: Gen manuals, 1997; Devlin, 1997; Raghunathan, 1996 reported that the use of CASE tool would significantly enhance the DSS development. The ARDSS was developed using Cool: Gen CASE tools, the use of CASE tools enable the ARDSS to carry the following advantages:

    a. Development cost savings;

    b. Developer productivity increases;

    c. Improvements in business processes;

    d. Higher levels of users' satisfaction;

    e. Flexible and high performance applications;

    f. Applications ease of use;

    g. High greater growth potentials;

    h. Supports most of the leading RDBMS and many OS environments;

    i. The ability to generate code indifferent;

    j. Comprehensive documentation;

    k. The link between the business objectives and the development activities.

Chapter summary

- The ARDSS consists of the following components:
    - Data management;
    - Knowledge;
    - User interface;
    - The users.
- The ARDSS capable of taking the following decisions:

- Accept or reject a new applicant;
- Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records;
- Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major;
- Hold the applicant until the following term/year;
- Accept or reject the applicant who is transferred from another educational institution;
- Predict a student's performance based on the students' history we keep;
- Classifying students into similar groups;
- Predict a student's performance based on the group that he/she belongs to;
- Set the student status to "On probation";
- Make relationships between students' performance and academic departments;
- Decide on Student abandonment.

- The ARDSS was successfully tested on four levels:
  - The consistency check level provided by Cool: Gen CASE tools;
  - Professionals level;
  - Users level;
  - Rules level.
- The ARDSS has the following limitations:
  - Restricted to the knowledge stored in its knowledge-base;
  - Able to take only twelve decisions;
  - Is an environment-specific system;
  - Designed for the Egyptian Universities to be used by certain managers;
  - The discovered knowledge is based on records drawn from the AASTMT students' DB;
  - The ARDSS has not been installed
- The ARDSS has the following management implications:
  - Gaining competitive advantages;
  - Managers are more committed and informed;
  - Better-served customers;
  - The benefits of using CASE tools.

Chapter eight

Conclusions and Recommendations

8-1 Conclusions

8-1-1 Chapter two

In chapter two many issued about DSS were discussed and evaluated. Foremost amongst which are the following:

- The differences between data, information, and knowledge have been set at the beginning of the chapter. *Data* is raw facts about things, events, activities, which are classified and recorded but not organised to convey a specific meaning. *Information* is processed data that is characterized as being; clear, provided to the relevant person at the right time, error-free & omission-free, and submitted to the decision maker frequently. *Knowledge* is a combination of experience, accumulated learning, and information that have been organised and analyzed to be understandable and applicable to a specific decision situation. The reason for drawing the differences between the three terms is the focus of this research. This research is focused on how to deliver information and knowledge to business managers to better understand their business problems and hence to improve the quality of the decisions they make.
- Different definitions for information systems have been analyzed and compared. The comparison revealed that Laudon and Laudon (2000) definition to IS is the most comprehensive one and therefore adopted in this thesis. The other definitions introduced for Alter (1992), Corr (1995), and Rowley (1996) focused on certain aspects of IS rather than being comprehensive. Laudon and Laudon defines an information system as a combination of work practices, information, people, and information technologies organised to accomplish organisational goals.
- There are different types of IS each has its own definition, characteristics, and usage. TPS record and collects data about the daily transactions taking place in any organisation; they are short-term in nature and have no decision capability. MIS provide the information required for managing the organisation. MIS require TPS as a prerequisite. ES are designed to help managers make better decisions in certain areas.

They are interactive CBIS that respond to questions and give recommendations. OAS refer to all the CBIS associated with general office work applications. ANN attempt to tease out meaningful patters from vast amounts of data and can recognize too many patterns. EIS/ESS are highly interactive systems providing managers and executives with flexible access to information for monitoring operating results and general business conditions. Executives use EIS without any aid from intermediaries.

- Managers use information systems in a variety of decision making situations. For structured decisions managers use TPS and MIS. For unstructured decisions, DSS, ES and ANN are used. EIS are special type of information systems that support unstructured decisions.
- The strategic role of IS involves the development of products and services that give the organisation strategic advantages over the strategic forces in the market. A strategic information system (SIS) is any IS that can help organisations achieve competitive advantages including: TPS, MIS, EIS, or DSS.
- *Discussion of the study's first objective No. 1 "Investigate and critically evaluate the current DSS practices"*. Discussion of this research objective revealed that past studies on DSS have made significant contributions, however, new contributions are required to address their shortcomings. The past DSS research has focused on many issues, but no integrated approach has been found because each study has tried to narrow the different aspects of the DSS. For example, in the 70's the DSS definitions (Scott-Morton; Little) focused on data processing, and model. During the 80's the definitions (Alter; Bonczek; Keen; Sprague and Carlson; Benett) focused on CBIS, effectiveness and the knowledge component appeared as part of the DSS. Whilst during the 90's the DSS definitions (Stevens; Corr; Reynolds; O'Brien, Marakas) focused on the models and problem structureness with more attention given to the knowledge component. Recently (Long and Long) definitions frequired the DSS to be interactive and user-friendly. However, no definition has been found to include all the following aspects:

the type of data used, the management level, the DSS effect, effectiveness of the DSS, type of knowledge targeted by the DSS. From this analysis it was evident that none of the definitions found was comprehensive, and this conclusion led to the need for a new DSS definition and methodology that are to be comprehensive.

- DSS are characterized by:
    a. They can be applied to support operational, tactical, and strategic level problems;
    b. Assisting managers to make repetitive decisions;
    c. Helping managers to evaluate options and choose the best one;
    d. Working within a short-term frame;
    e. Handling complex problems;
    f. The need for an interaction between the decision-maker and the DSS;
    g. Being developed by non-data processing (DP) professionals;
    h. Focusing on the flexibility of decision making;
    i. Providing information to support certain decision area;
- The components of a DSS are data management subsystem, model management subsystem, knowledge management subsystem, user interface subsystem and the system user.
- DSS can be classified according to many factors by many authors including:
    a. Donovan and Madnick 1977 classified DSS by the type of problems into institutional and Ad hoc DSS;
    b. Alter 1980 classified DSS by the degree of action implication of system outputs into data oriented, model oriented, and data and model oriented DSS;
    c. Bonczeck et al. 1980 classified DSS by the degree of non-procedurality into non-procedural languages based and procedural languages based DSS;

d. Hackathorn and Keen 1981 classified DSS by the type of support into personal, group, and organisational DSS;

e. Holsapple and Whinston 1996 classified DSS by its orientation into text, database, spreadsheet, solver, rule, and compund DSS.

- Decisions are taken by either individual manager or group of them, when the decision is taken by a group of managers this is what is called GDSS.
- DSS can be developed using a number of approaches including the following:

a. SDLC. The SDLC contains eight development steps, however, not all DSS go through all of these steps (Meador et al., 1998; Keen and Scott Morton, 1978). These steps are: planning, research, system analysis and conceptual design, design, construction, implementation and user training, maintenance, and adaptation.;

b. Prototyping. Starts by select an appropriate subproblem to be built then develop a small usable system for the decision maker to evaluate and then refine and modify the system in cycles;

c. End-user. It is the development of CBIS by people outside the formal information systems area e.g. managers, financial analysts, engineers, and lawyers. They build DSS to support their work and enhance their productivity.

8-1-2 Chapter three

In chapter three the the data warehouse was introduced and it was critically evaluated. The chapter revealed the following:

- Onder and Nash, 1999; Srivastava and Chen 1999; Barquin, 1997; Berson and Smith, 1997; Paller, 1997 emphasized that the need for a data warehousing technology has been driven to meet the managements' information needs including:

a. Decisions need to be taken quickly and correctly using all the available data;
b. Users are not computer professionals so they need all the relevant data concerning a specific business problem to be stored in one place;
c. The amount of data concerning a specific business problem is increasing;
d. It is becoming increasingly important to be able to obtain a comprehensive and integrated view of the enterprise for the purpose of making decisions;
e. Decisions sometimes require historical analysis;
f. Increasingly businesses are working closer together and are able to share and exchange data in what is known as a strategic alliance;
g. Identifying trends in the business.

- Different DW definitions were illustrated and investigated. However, each of the definitions tried to define the DW taking into consideration certain point of view. Some definitions are technical (Berson and Smith, 1997), others are about the use of DW (Berson, 1996), the features (Inmon and Hackatorn, 1994), or the goals (Kimball, 1996). However, no definition has been found to be comprehensive covering the data sources, front-end, and the purpose of the DW in a business context. This has led to the development of a new DW definition.
- *The DW proposed definition*: "A DW is a group of data extracted from different sources; internal, external, historical, and personal data archived in one or more data stores. The purpose of constructing a DW is to provide the DSS and the decision maker with the necessary data, which when transformed into information, will provide a better understanding of the business problem."
- The DW benefits were also clarified.
    a. increses the decision maker's productivity by providing accessible data in a ready to use format;

b. more cost-effective decision making process by separating the query processing from the operational databases;
c. enhancing asset and liability management by providing the overall picture of the enterprise purchasing and inventory transactions;
d. supporting the corporate strategy that positions the clients at the center of all operations which could not be achieved without a DW;
e. reduces redundant processing, support, and software to enhance DSS applications;
f. enhancing the work process, which also affects the success of business process reengineering;
g. improve customer service;
h. organisations will be able to exceed competitor capabilities and achieve competitive advantages.

- The differences between DW, data marts, and enterprise data warehouses (EDW) were investigated deeply. A smaller local data warehouse that is classified by subject is called a data mart. An EDW is a planned, integrated, managed store of relevant corporate data optimized for analysis, query, and reporting functions. An EDW contains large number of fields and millions of data records about the entire organisation.
- The differences between the ODS databases and DW database design were investigated. The ODS databases provide the DW with a source of data, however, they lack the functions required to perform efficient analysis and produce reliable results that decision makers really need. Moreover, the contents of the DW are relatively stable whilst the contents of the ODS change as each transaction is initiated.
- The DW database can be built using the relational model or star schema structure. The star schema structure is the best for DW design.

- The star schema structure/MDDM captures the measurements of importance to the business and the parameters by which the business measurements are broken down. The measurements are referred to as facts, whilst the parameters by which a measurement can be viewed are called dimensions.
- The data warehouse application is a client-server application, and preferably multi-tiered. Analysis revealed that the multi-tiered is less complex, highly secured, and can support the internet applications and can handle heterogeneous DB environments.
- A DW development strategy is required at the early stages of the project; this strategy is described in a *DW development strategy document*. A DW development strategy document contains the following items: the DW project objectives, duration of the project, the approach used, a DW rollouts plan, users, DW architecture, data sources, and the DW updating policy.
- Data mining is a useful technique for extracting information and knowledge from the DW. Data mining is the process of finding hidden knowledge and unknown facts and trends in data. Data mining is a step of the knowledge discovery in database (KDD) process.

8-1-3 Chapter four

In chapter four the KDD process was defined and analyzed. Among the issues discussed in chapter four are:

- There are four different types of knowledge; shallow, multi-dimensional, hidden, and deep knowledge. Shallow knowledge can be discovered using SQL, multi-dimensional knowledge can be discovered using OLAP and visualization, hidden knowledge can be discovered using group of data mining techniques, and the deep knowledge can be discovered by employing the entire KDD process.
- KDD is the process of finding hidden knowledge, patterns and unknown facts from the data sets. Data mining is a step in the KDD process.

- KDD is a multi-disciplinary field, and the KDD process has been used successfully in many disciplines.
- The KDD process is interactive, iterative, and involves a great deal of user-interference. Brachman and Anand in 1996, defined the practical view of the KDD process as follows:
    a. Developing an understanding of the application domain;
    b. Creating a target data set;
    c. Data cleaning and preprocessing;
    d. Data reduction and projection;
    e. Choosing the data mining goal and task;
    f. Choosing the data mining technique(s) or algorithm(s);
    g. Data mining;
    h. Interpreting the information gained by the mining techniques;
    i. Consolidating the discovered knowledge.
- From the work of Brachman and Anand (1996):
    a. KDD is an entire process that should be applied from the application domain identification step until the evaluation of the discovered knowledge;
    b. The discovered knowledge should be then utilized in a suitable front-end tool like EIS or DSS;
    c. A DW will enhance the KDD results;
    d. organisations would get the maximum benefits from their business applications if they include these components (i.e. DSS, DW, and KDD).
- Various data mining techniques were discussed including:
    a. SQL. Query tools are used to extract data that matches search criteria or to represent this data in a way that the user finds easier to handle or interpret. By applying simple SQL users can obtain a wealth of information;

b. Visualization. Visualization techniques depend strongly on the human side of the analysis (Berson, 1996). Data visualization is emerging as a technology that may allow organizations to process amounts of data and present it in a usable format;

c. OLAP. The key driver for the development of OLAP is to enable the multi-dimensional analysis (Pyle, 1999). Although all the required information can be formulated using relational database and accessed via SQL, the two dimensional relational model of data and SQL have some serious limitations for investigating complex real world problems. Also slow response time and SQL functionality are a source of problems (Berson, 1996). OLAP is a continuous and iterative process; an analyst can drill down to see much more details and then he can obtain answers to complex questions.;

d. Association rules. According to Wijsen, 2001; Liu, et al., 2000; Aas, et al., 1999; Berson and Smith, 1997; Adriaans and Zantinge, 1996; Agrawal, et al., 1996, the interest in discovering association rules from large relational tables has been increased recently. Association rules are focused on finding relationships (i.e. associations) between a certain attribute (i.e. target attribute) that the user is interested in, and the remaining attributes in a relational table. The strength of association rules is that they can efficiently discover a complete set of associations that meet the user's requirements. However, there is no single algorithm that will automatically give the users everything of interest in the database.;

e. Cluster analysis. Clustering is basically classifying unclassified data (Gordon, 1981; Everitt, 1980). The data to be classified consists of a set of items (sometimes referred to as objects, fields, or records). Each item is described by a set of characteristics called variables (sometimes referred to as attributes). The target of clustering is to classify the items in the data set

into a number of groups (sometimes referred to as classes, or clusters), such that objects within one group have similarities with one another. Where the number of items is n, the maximum number of groups should be n-1 (i.e. the number of expected groups varies from 1 to n-1);

f. Decision trees. A decision tree is a predictive model that provides a means of visualizing complex decision problems where the questions can be posed in sequence;

g. GA. The term is a combination of both biology and computer disciplines, and sometimes referred to as simulated evolution. Berson and Smith (1997) said that GA refer to these simulated evolutionary systems, but more precisely these are the algorithms that dictate how populations of organisms should be formed, evaluated and modified;

h. ANN. An ANN is a computer programs that implements complex pattern detection and machine learning algorithms to build predictive models from large database(s). In order for the ANN to detect patterns in the data sets, it should learn to detect these patterns and make predictions, in the same way a human does. ANN are widely used in many business applications;

i. Probabilistic graphical dependency technique. These models specify the probabilistic dependencies, which underlie a particular model using a graphical structure. The model specifies which variables are dependent on each other. These models are used for categorical or discrete-valued variables, however, some extensions also allow for the use of real-valued variables (Fayyad et al., 1996).

- Each data mining technique has its own features and characteristics, hence not all the data mining techniques are applied to a certain application. Rather, each application applies only one or more data mining techniques that best fit with the data to be analyzed and the objective of the entire KDD process.

- When applied, each data mining technique should have a goal and a task. The goals of the data mining techniques are prediction and description, whilst the data mining tasks are clustering, classification, summarization, dependency, regression, and change detection.
- The KDD process were applied to 1600 records drawn from the AASTMT application records. The objective of this application was to put the KDD into a practical context. Various activities were performed in this application including: data cleaning, coding, enrichment, and some mining techniques (SQL, visualization, OLAP, clustering, and decision trees).
- The spread of the KDD process especially the data mining techniques will be enhanced many folds if used with the data warehouse. Moreover, the DW need a front-end tool (i.e. DSS), and this is the basic idea of this research, combining these three ingredients together.

8-1-4 Chapter five

In chapter five the three components of the proposed DSS methodology were linked together, and importantly a new DSS definition, and a new DSS development methodology were developed:

- In order to produce effective decision support systems that are able to enhance the quality of the decision making process research efforts have concentrated on introducing different combinations of three components (i.e. DSS, DW, and KDD), whilst other research efforts have ascertained the importance of linking them together. Moreover, this research has found that there have been very little efforts made to integrate these three components in certain application areas or to explain how these components work together and what tools and mechanisms used.
- The research efforts can be classified into two groups: the first group includes those who have tried to show the importance of linking two components together (Inmon and

Hackathorn, 1994; Mattison, 1997; Turban and Aronson, 1998; Humphries, et al., 1999). The second group includes those who have assured the importance of linking the three in future research (Adriaans and Zantinge, 1996; Han, et al., 1999; Cooper, et al., 2000).

- Due to the shortcomings of the traditional DSS definitions evaluated in chapter two, and to respond to the new methodology which combines DSS, DW, and KDD a new DSS definition will be introduced: "A DSS is a computer-based information system that deals with semi-structured and unstructured problems facing managers at all management levels. The DSS goal is to enhance the decision quality and the manager effectiveness. To do so, the DSS integrates itself to the strategic data store which is the data warehouse (DW), and to the knowledge discovery in database (KDD) process that will find the deep knowledge and hidden patterns in the DW and present them to the DSS user."
- The research objective No.2 "Develop a new DSS methodology" was evaluated in this chapter and a new DSS proposed DSS methodology was also introduced. The methodology consists of four modules. In Module 0 the needs' analysis is performed, the module uses a questionnaire for collecting the user requirements and Cool: Gen CASE tools planning and analysis phases to maintain and implement the requirements. In Module 1 the data warehouse will be built, the module uses MS-SQL Server, star schema structure, and Crystal reports. In Module 2 the KDD process is to be applied, the module uses MS-SQL Server, Cool:Gen CASE tools analysis and design phases, and utilizing the following data mining techniques: SQL, visualization, and clustering analysis. In Module 3 the DSS will be built, the module uses Cool: Gen CASE tools design and generation phases.
- Justification of the chosen tools and techniques proposed by the new DSS methodology was investigated deeply. For example, Cool: Gen 5.0 by Sterling is the CASE tool employed in this research to capture the information needs at the highest possible level

of abstraction and transform them into executable application systems. The choice of Cool: Gen CASE tools is based on the following backgrounds: cost savings achieved by using the Cool: Gen CASE tool, developer productivity increases by up to 300%, dramatic improvements in business processes, higher levels of customer satisfaction, extraordinary flexible and high performance applications, accelerated systems development, and greater growth potential. Also, MS-SQL Server was chosen to be used in the development of the data warehouse because it supports both the relational and star schema structure models efficiently. Moreover, the data mining techniques were critically evaluated and based on this evaluation SQL, visualization, and clustering analysis were chosen.

- Different proximity measures were critically evaluated including *Gower, Jaccard, City Block, Euclidean, Modified Euclidean,* and *Canberra.* The evaluation obtained the following results:

    a. The research results are consistent with the previous related work in two findings. Firstly previous related work indicated that Gower is the most suitable measure where the data types are mixture. Secondly previous work also did not recommend Jaccard where the data types are mixture. When applying Gower and Jaccard similar results were obtained; Gower was able to detect similarities and dissimilarities between the different record groups, whilst Jaccard obtained an extreme value (i.e. 0 or 1). This is because Jaccard is based on the number of similarities and dissimilarities and it completely discards the attribute values, it is possible that its value could be 1 or 0;

    b. Moreover, City block generated a value of 0.5 for the first group that includes two very similar records. That makes it incomparable with the 0.99 value produced by Gower or the 0.92 produced by modified Euclidean for the same group. In addition to that, according to City block the similarity of

the records in the second (i.e. dissimilar records) is .017 which is higher than the similarity of the records in the last group .01 (fairly dissimilar records);

c. Although Euclidean metric measure did not perform consistently well as Gower and modified Euclidean, it was decided to be retained and used with the clustering techniques because of the following reasons: Euclidean was recommended to be used with many clustering techniques (e.g. Nearest Neighbour, Furthest Neighbour, Ward's) by many scholars (Wishart, 2000; Gordon, 1981; Everitt, 1980), also the results of the Euclidean measure are good because it was able to distinguish between similar and dissimilar records;

d. Another finding of this research is the results obtained by the modified Euclidean and Canberra metric measures (after transforming their results from distance metric to equivalent similarity coefficients). Previous work shows that little analysis has been carried out to apply the modified Euclidean and Canberra metric measures to a mixture of data types. However, the results of this research indicate that they are in close agreement with the results obtained by Gower's similarity measure.

- Various clustering techniques were studied including *hierarchical methods, optimization, density,* and *clumping*. The results obtained revealed the following:

a. The analyst must know that there are many clustering techniques available each of which has its own assumptions and gives different results if applied to the same data sets, the decision on choosing the clustering technique should be made in the light of the advantages and disadvantages of the chosen technique and the type of data to be classified;

b. Deciding on the number of clusters is a problem that is common to most of the clustering techniques. There are a lot of research efforts to handle this

problem, however, there is no satisfactory solution for determining the optimum number of clusters (Aas, 1999; Gordon, 1981; Everitt, 1980). There is no optimum number of clusters, it largely depends on the problem being resolved;

c. Although the concept of hierarchical techniques was developed in biology, this group of techniques is now used in many areas. One advantage is that the question on the optimum number of clusters does not arise since the researcher is interested in the complete hierarchy;

d. The biggest disadvantage associated with these techniques is their inability to reallocate items, which might be misclassified at early stages (Everitt, 1980). However, others (Jardine and Sibson, 1968) said that the Nearest Neighbour (NN) has the greatest mathematical appeal amongst, and would generate suitable results for most application areas;

e. The NN and FN techniques have the problem of *chaining* (Wishart, 1998; Everitt, 1980). Chaining means that the method tries to accumulate the new records/case on existing cluster(s) rather than creating new clusters. As a result of that, the number of records/cases in each cluster is highly affected by that problem, that is, few clusters retain the majority of records whilst the remaining clusters have quite little number of records;

f. The optimization techniques suffer from a number of problems. The techniques are transformation dependent; that is different results would be obtained from applying the same technique to the same data set. However, the advantages of using optimization techniques are the ability to reallocate misclassified item in further stages, and these techniques also do not assume that all clusters are hyper spherical (i.e. have the same shape). The most serious problem with the optimization techniques is the large amount of

computation power they require, which in turn makes them irrelevant for the very large data sets;

g. The density solutions suffer from the problem of sub-optimal solutions; they might be more than one solution for the data sets (i.e. maximum likelihood). TAXMAP also suffer from the problem of containing various parameters that control the technique and arbitrary chosen by the investigator (Everitt, 1980);

h. The Clumping techniques rather than their unsuitability for the data sets, they suffer from the problem of optimization techniques that is the computation power they need.

- Based on the analysis and evaluation of the various proximity measures and clustering techniques, it was decided that Ward's will be used base on a Euclidean metric proximity measure. Ward's was less affected by chaining than NN and FN. Ward's also obtained a logical and consistent results because similar students came in a consequent order and dissimilar students did not come in a consequent order.

8-1-5 Chapter six

In this chapter the users' requirements were identified and analyzed. The ARDSSQ was used to identify the users' requirements, whilst frequency tables, Chi squared statistics, and canonical correlations were used to analyze the responses. Among the important issues discussed in this chapter are:

- The research problem was defined as the Admission and Registration function in Universities.
- The research population was defined as the Egyptian Universities both Government and Private.
- The response base was identified as Deans, Associate Deans, Admission Officers, Registrars, and Others.

- Searching the literature revealed that no research questionnaire was found relevant. Hence, a new research questionnaire (ARDSSQ) was developed and validated. The ARDSSQ consists of seven constructs, each of which investigates one research objective.
- The population was contacted and asked to participate and then the ARDSSQ was sent to those who agreed to participate.
- The response rate attained was **92.3 %** on the University level, and **24.9 %** on the respondent level.
- Each of the questionnaire constructs will be analyzed in terms of these three dimensions: University type, Respondent position, and Whether the University uses a CBIS or not. The reason for using these three dimensions is due to the nature of the population, the response base, and the expected effects of using CBIS on the answers. And also to identify areas of commonality and discrepancy between the major segmentations identified within the population.
- Discussion of the research objective No. 3-1 "The managers' perspectives towards computers and their current admission and registration information systems":
    a. The managers' perspectives towards computers and their current admission and registration information systems is affected by these two dimensions; the University type and the managers' position;
    b. The use of CBIS does not affect the managers' perspectives towards computers and their current admission and registration information systems;
    c. The percentage of Private Universities that use CBIS (**92 %**) is greater than the percentage of Government Universities that use CBIS (**35 %**);
    d. The DW component is expected to enhance the decision quality.
- Discussion of the research objective No. 3-2 "Features of these information systems":

a. The features of the current Admission and Registration IS in the Egyptian Universities are not affected by any of these dimensions; the University type, the respondent position, and the use of CBIS;

b. The Admission and Registration information systems in the Egyptian Universities have the following features: Printing reports that describe students' records feature, Electronic stores of students' data;

c. The Admission and Registration information systems in the Egyptian Universities do not have the following features: Predicting the new applicants' performance, and Predicting the current-students' performance.

- Discussion of the research objective No. 3-3 "Functions of these information systems:

a. The functions of the current Admission and Registration information systems are affected by these two dimensions; the University type and the manager's position;

b. The functions of the current Admission and Registration information systems are not affected by the use of CBIS dimension;

c. Different management levels require different information needs;

d. The Admission and Registration information systems in the Egyptian Universities have the following functions: Student description reports, General statistics, Classifying students into similar groups, Using the historical data to describe the Students' history (*only in the Private Universities*);

e. The Admission and Registration information systems in the Egyptian Universities have do not have the following functions: Student performance prediction, and Finding relationships between a student's data fields.

- Discussion of the research objective No. 4-1 "The managers' perspectives towards the role of computers and the ideal admission and registration information systems":

a. The managers' perspectives towards the role of computers and the ideal Admission and Registration information system is affected by both the University type and the manager's position dimensions;

b. The use of CBIS does not affect the managers' perspectives towards the role of computers and the ideal Admission and Registration information system;

c. Respondents from the Private Universities have a better understanding to the managers' perspectives towards the role of computers and the ideal Admission and Registration information system;

d. There is a need for the DSS to be developed for the Admission and Registration functions. The results of Q.21 "*The admission and registration information system should be able to help managers take decisions*" showed that 98% of the Private Universities, and 100% of the Government Universities think that their Admission and Registration information system should be able to help managers take decisions.

- Discussion of the research objective No. 4-2 "The decisions that are expected to be taken by DSS":

a. The decisions that are expected to be taken by the Admission and Registration DSS are affected by these dimensions; the University type, the manager's position, and the use of CBIS;

b. The Admission and Registration DSS should be able to take the following decisions:

1. Accept or reject a new applicant;
2. Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records;

3. Predict the new applicants that will join the faculty/college/institute this term/year based on government statistics on secondary school students;
4. Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records besides other records like the government statistics;
5. Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major;
6. Hold the applicant until the following term/year;
7. Accept or reject the applicant who is transferred from another educational institution;
8. Accept or reject the applicant who is transferred from another educational institution based on our transfer history records;
9. Predict a student's performance based on the students' history we keep;
10. Predict a course's results based on the courses' history we keep;
11. Classifying students into similar groups;
12. Predict a student's performance based on the group that he/she belongs to;
13. Set the student status to "On probation";
14. Predict the "On probation" students based on the students' history we keep;
15. Make relationships between students' performance and academic departments;
16. Forecast course booking;
17. Decide on Student abandonment.

- Discussion of the research objective No. 4-3 "4-3 DSS functions":

a. The ideal Admission and Registration information system functions are affected by both the University type and the manager's position dimensions;
b. The ideal Admission and Registration information system functions are NOT affected by the use of CBIS dimension;
c. The ideal Admission and Registration information system should have the following functions:
    1. Predict new applicants' performance (*only for the Government Universities*);
    2. Predict current students' performance;
    3. Producing student description reports;
    4. Provide general statistics;
    5. Student classification into groups;
    6. Using historical data;
    7. Being able to use external data;
    8. Finding relationships between students' data fields;
    9. Gives the user the ability to create ad hoc reports.

- Discussion of the research objective No. 4-3 "4-4 DSS characteristics":
    a. The ideal Admission and Registration information system should have the following characteristics:
        1. Easy to use;
        2. Requires minimum training;
        3. User involvement in design of the system;
        4. Able to grow;
        5. Flexible;
        6. Integrated;
        7. Have E-mail facility;
        8. Web-accessible;

9. And cost effective.

- Statistical analysis using Chi-squared statistic revealed the following:
    a. There is a significant relationship between the University type and the use of CBIS by which the proportion of private Universities who are using CBIS differs from that proportion in the government ones;
    b. There is a significant relationship between the respondent position and the use of CBIS by which the proportion of Deans who are using CBIS differs from that proportion with regard to the Associate Deans, Registrars, Admission Officers, and Others;
    c. There is a significant relationship between the respondent position and acceptance to role of computers as data stores by which the proportion of Deans who believe that being a data store is the main role of computers differs from that proportion as for Associate Deans, Registrars, and Others;
    d. There is no relationship between the respondent positions and the role of computers as being decision makers;
    e. There is a significant relationship between the respondent position and the ownership of a PC on his desk by which the proportion of Deans who have a PC's on desks differ from that proportion for the Associate Deans, Registrars, Admission Officers (Admission Officer and Associate Deans are the same), and Others.

- Statistical analysis using Canonical Correlations revealed the following:
    a. The Canonical function found is:

$$-0.035 * Y\_1 + 0.012 * Y\_2 - 1.016 * X\_3 = 0.174 * construct\_1 + 0.752 * construct\_2 + 0.105 * construct\_3 + 0.018 * construct\_4 - 0.009 * construct\_5 - 0.035 * construct\_6 + 0.004 * construct\_7$$

    b. Where the magnitude of the variable represents its contribution to the variate it belongs to. Variables of opposite signs represent inverse

relationships to each other's. That is, among the independents' variate the CBIS use (X_3) accounts for the highest effect and works on the same direction as the University type (Y_1) and both are opposite to the Manager's position (Y_2) which has the least effect on the variate. Also the second, first and third constructs (in order) have the highest effect on the variate of the dependent variables. Among the dependent variables only the fifth and sixth constructs move in the opposite direction to the remaining constructs. The reason why the first three constructs have the highest relationship magnitudes is because they representing the current Managers' perspective towards CBIS, the current Admission and Registration IS features, and the current Admission and Registration IS functions which are highly affected by the three independents, whilst the remaining constructs are about ideal Managers' perspectives and ideal DSS decisions, functions, and characteristics where the three dimensions have little impact.

8-1-6 Chapter seven

In this chapater the proposed DSS methodology was used to develop the ARDSS. The chapter discussed the following:

- The research objective No. 5 "Use the proposed methodology to develop the required Admission and Registration DSS" was investigated. That is, the proposed DSS methodology will be applied to the ARDSS development. The objective has been met by implementing the methodology's four modules.
- Module 0: Needs' Analysis. This module was accomplished in four phases as follows:
    a. the first phase is the development and validation of a new research questionnaire that is used to define the current Admission and Registration information systems in the Egyptian Universities and to explore the requirements that are not satisfied by these current systems;

b. the second phase the questionnaire was used to collect data from the Admission and Registration Managers in the Egyptian Universities;

c. the third phase, the Managers' information needs that are required to be satisfied (Refer to chapter six for more details) have been identified;

d. the last phase, Cool: Gen CASE tools Planning and Analysis phases were utilized to start the development.

- Module 1: Building the data warehouse. In this module the University DW was designed and implemented on MS SQL Server. In this module the research objective No. 2-2 "Designing the DW" was achieved by the undertaking the following five steps:

  a. Study and evaluate the data sources;

  b. Establish the source-to-target fields' matrix as a design validation tool;

  c. Build and the DW Star Schema design using MS SQL Server;

  d. The Updating strategy of the DW;

  e. Design the Managers' reports using Crystal Reports.

- Module 2: Knowledge from the KDD process. In this module the KDD process was applied to 1800 records. SQL, Visualization, and Clustering analysis techniques have been used as data mining techniques. The techniques have been applied for the following reasons:

  a. Describing and representing the sample;

  b. Finding the knowledge which will be stored in the DSS knowledge-base;

  c. Creating the managers' reports from the DW.

- Module 3: Building the ARDSS. In this module the following components have been identified:

  a. *Data management component*. In this component the ARDSS DB has been created in Cool: Gen CASE tools and being transferred to MS-SQL Server (Refer to Appendix (F) for details);

b. *Knowledge management component*. The results of applying the knowledge discovery techniques were stored using the design tool provided by Cool: Gen CASE tools. As new data is added to the DW, the knowledge discovery techniques run again and new knowledge could be found, which could affect the knowledge base by adding new rules, changing or modifying or deleting existing rules;

c. *User Interface*. The ARDSS adopts a GUI environment;

d. *Users*. The users of the ARDSS have been identified Dean, Associate Deans, Registrars, Admission Officers, and Others.

- The ARDSS capable of taking the following decisions:

1. Accept or reject a new applicant;
2. Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records;
3. Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major;
4. Hold the applicant until the following term/year;
5. Accept or reject the applicant who is transferred from another educational institution;
6. Predict a student's performance based on the students' history we keep;
7. Classifying students into similar groups;
8. Predict a student's performance based on the group that he/she belongs to;
9. Set the student status to "On probation";
10. Make relationships between students' performance and academic departments;
11. Decide on Student abandonment.

- Testing the ARDSS. The ARDSS has passed successfully four level of testing: consistency check level, professionals' level, user level, and model level (using 200 records).

- The ARDSS limitation. Following are the limitations of the ARDSS:
    a. restricted to the knowledge stored in its knowledge-base;
    b. The ARDSS is able to take only eleven decisions, which is turn means that not all of the Admission and Registration related decisions are incorporated into the system;
    c. The ARDSS is an environment-specific system; that is it requires a Client/Server environment which has MS SQL Server 6.5 RDBMS running on a Windows NT 4 OS, Crystal reports 4.6, CLUSTAN graphics 5.0, and a Windows 95 or 98 on a Pentium machine;
    d. The ARDSS is designed for the Egyptian Universities to be used by Deans, Associate Deans, Registrars, Admission Officers, and Others;
    e. The discovered knowledge (Refer to chapter seven section 7-5-2) is based on records drawn from the AASTMT students' DB. Although other Universities' managers including both Government and Private found the majority of the knowledge base relevant and acceptable, this knowledge can only be used for decision making at the AASTMT, and if any other University will use the ARDSS records from this University need to be included in the KDD process;
    f. The ARDSS has not been installed (Refer to chapter seven section 7-6-3).
- The management implications of the ARDSS. The following management implications have been identified:
    a. Gaining competitive advantages;
    b. Managers are more committed and informed;
    c. Better-served customers (i.e. students);
    d. The benefits of using CASE tools.

8-2 Recommendations

8-2-1 For the Egyptian Universities

- The Ministry of Higher Education and the Egyptian Supreme Council of Universities are currently setting the standards and regulations which both University types (i.e. Government and Private) have to follow. It is recommended that these two bodies set the same Admission and Registration standards and regulations for both University types. The researcher has found some difficulties in some areas and situations whenever a comparison is to be made between the two University types. Examples are the following:

    a. The respondents. In some Government Universities the Admission and Registration decision makers are: *Deans*, *Associate Deans*, *Admission Officers*, and *Registrars*. However, in other Government Universities there are no *Admission Officers or Registrars*, the two positions have been replaced with a position called *Director*. On the other hand, in the Private Universities the decision makers are *Deans*, *Associate Deans*, *Admission Officers*, and *Registrars*;

    b. Position responsibilities. There is also a difference in the definition of the *Registrar* as a position in both University types. In Government Universities the registrar is a college/school level position, whilst in Private Universities it is a University level position. The difference in responsibilities would affect their information needs;

    c. The academic year definition. For some of the Government Universities the academic year starts in September and ends in July, whilst other Government and all Private Universities have a semester-based academic year; September, February and Summer semesters. This also affect the design of the Admission and Registration Information System that they may

use. For example the Admission procedure happens once a year in some of the Government Universities and more in others;

d. The grading system. Some Government Universities follow a descriptive grading system (Excellent, Very Good, Good, Satisfactory, and Poor), whilst others follow a GPA grading system (scale of 4). All of the Private Universities have the same grading scale (GPA);

e. The DSS unit. Some Government Universities have recently established a University-level DSS unit. The mission of this unit is not yet clear because until the data collection time (June-December 2000) the DSS units did not have a particular IS strategy;

- It is advisable that both University types use the ARDSS proposed by this thesis[1] because of the following:

  a. The current Admission and Registration systems are incapable of providing the managers with their information needs. For example **69%** of respondents in the Private Universities and **73%** in the Government Universities reported that their systems can not predict the performance of their current students. However, **69%** of respondents in the Private Universities and **78%** in the Government Universities reported that they need their systems to predict the performance of current students. The ARDSS has the performance prediction capability;

  b. The ARDSS fulfills an important design principle "User-involvement during the system development process", on which **86%** of respondents in the Private Universities and **83%** of respondents in the Government Universities reported positively;

[1] Applying the ARDSS in any University requires modifications to the system to respond to the University-specific regulations and structure.

c. The ARDSS is enhanced with a data warehousing component which gives the managers access to some reports which are based on 10 year time span (or more if data is available), and are combining fields/attributes from many tables/entities without affecting the performance of the operational DB. This data warehouse component adds a strategic value to the use of the ARDSS;

d. The ARDSS is built utilizing knowledge discovery in database techniques (i.e. SQL, visualization, and clustering analysis) which give the system the capability to reach different types of knowledge (shallow, multi-dimensional, hidden, and deep knowledge);

e. The use of CASE tools in the development of the ARDSS added the following advantages to the system:

   i. Flexibility. The system can easily be changed to respond to new business needs. Also any business object(s) can be included into and/or removed from the system easily;

   ii. Scalability. The ARDSS is scalable and can be generated to be applied in different hardware and software environments easily;

   iii. Integrity. The ARDSS can be easily integrated with other information systems;

   iv. Documentation. The ARDSS is well documented and this makes the future modification and/ maintenance of the system easy even if the designer is not present;

   v. Business needs. Business needs and objectives are the main driver of the ARDSS;

   vi. Interface. The ARDSS is based on an easy to understand and deal with GUI;

- It is also advisable that the Admission and Registration managers in Government Universities use CBIS to perform their functions. The results of the ARDSSQ showed that **72%** of them have PC's, whilst only **35%** of them use CBIS. The justification for the difference between the two percentages could be that they use their PC's in secretarial and administrative work (i.e. word processing and/or web and e-mail access) spend more on technological equipments (i.e. PC's for Admission and Registration senior managers). When a comparison is held between the Government and Private Universities, the percentages are **67%** and **92%** respectively. This means that the majority (**92%**) of the Admission and Registration managers in Private Universities who have PC's use CBIS to perform their functions;

- All Universities should train all senior managers on how to use and understand technology. The results of the ARDSSQ showed that **80%** of the managers in Government Universities still think that the main role of computers is an electronic data store. Whilst only **56%** of them think that the one of the main roles of computers is a decision maker. However, in the Private Universities these percentages are **48%** and **65%** respectively. Based upon the results we can say that the understanding to the roles of technology in Private Universiteis is better than their rivals in the Government Universities, but they both need training;

- The Egyptian Universities must pay attention to the wealth of information stored in their Admission and Registration TPS. From these systems lots of valuable knowledge can be extracted. For example the Admission and Registration managers are interested in reports that are based on ten years time (and sometimes more). However, they can not get these reports from their current systems. Moreover, most of the reports they need are not built on their current systems, as a result of that there has been always a time lag between requesting a report and getting it. Some of these Admission and

Registration managers were interviewed and they described some of the reports that need to be created, the reports have been created based on the data stored in the University data warehouse (Refer to Appendix **D** for the details of the University data warehouse and those reports);

- Acquiring new hardware decision at the Egyptian Universities should take into consideration the required specifications to develop and use DSS, data mining techniques, and data warehousing.

- Some of the Universities showed inadequate understanding to the role of research. The researcher has found some difficulties during the data collection phase because of this insufficient understanding. For example some of the Universities agreed to participate, however, they restricted their participation to no more than three questionnaires (e.g. MUST and SONGOR). Another Universities refused to participate because the researcher belongs to a Private University. Based upon these facts, it is recommended that the Universities in Egypt train their senior managers on to proper understanding to the role of research and it importance for societies.

8-2-2 For researchers

- Researchers can use the ARDSSQ and apply it to different educational systems in different countries (modifications need to be made to reflect any country's specific educational system environment);

- The new DSS methodology proposed by this thesis is recommended rather than the traditional development approaches.

8-2-3 For systems analysts and designers

- Cool: Gen CASE tools is a recommended DSS development approach because of the following:
    a. Cost savings achieved by using the Cool: Gen CASE tools;
    b. Dramatic improvements in business processes;
    c. Higher levels of customer satisfaction;
    d. Extraordinary flexible and high performance applications;
    e. Accelerated systems development;
    f. Applications ease of use, high greater growth potentials, and personalized solutions with built-in automated decision support;
    g. Supports most of the leading RDBMS (e.g. ORACLE, SYBASE, MS SQL, IBM DB Server), and many OS environments (e.g. UNIX, Windows NT, Windows 95/200, OS/390);
    h. The ability to generate code in different languages (e.g. C++, COBOL), which means that developers do not need to know a wide range of programming languages. They are only required to understand one simple English-like toolset, Cool: Gen;
    i. Users and developers of decision support systems implemented using the traditional approaches always having data availability, and data management problems.

- Building a DSS enhanced with KDD techniques utilizing a DW component is a team-work project. Preferably, the team should encompass people of different backgrounds: senior managers (i.e. expected users'), system developer(s), data warehouse specialist, IS expert, a specialist in data mining techniques, and a project coordinator;

- The issue of software and hardware compatibility should be carefuly addressed before the project starts. For example, CASE tools work in a Windows environment, communicate with specific DBMS (e.g. MS SQL Server, SYBASE, ORACLE, and IBM DB Server). As a result of that it should be identified at the beginning which software is required, what releases, are the releases compatible (i.e. can communicate with each others), do we have the hardware requirements to run the software. The answers to these questions are very important and could affect the functionality of the developed system.

8-3 Future work

- *Web accessibility.* Implementing the Admission and Registration DSS on the Web, so that senior managers can log to the Web site and access the system remotely;

- *Systems integration.* Integrating the Admission and Registration DSS with other systems. e.g. the integration with the financial system;

- *New dimensions.* Add the students' dimension to the system, so they are able to access the system and consult it. For example, they can log on and ask: which major to join, which courses to register in a certain semester, what will be their expected grade point average;

- *Knowledge base comparison.* The knowledge base of the ARDSS is based on records drawn from a private University students' DB. Further research could be undertaken to build the knowledge base based on records drawn from a government University students' DB and then compare the results;

- *New subject areas.* The scope of the Admission and Registration DSS could be enlarged by including new subject areas into the analysis phase and hence increase the number of decisions that the system is able to make. E.g. staff members, and scheduling (i.e. time table preparation);

- *The Data Warehouse.* Implement the DW on the University level and devote a separate DW server to handle the huge amount of data. The DW implementation should be developed in the new release (7.0 or 2000) of MS SQL Server (or any other DB Server taken into consideration the software and hardware compatibility), because new

releases include more data warehousing features e.g. star schema structure, report generation tools, and OLAP data cube structures;

- *Inter Universities Web-Based Data Warehouse.* Develop an inter Universities DW, from which all member Universities can extract useful information and helping them enhancing their business understanding and accordingly the decision quality;

- *Different data mining techniques.* The Admission and Registration DSS depends on SQL, visualization, and clustering analysis techniques, however, different techniques could be studied and applied to the system where more data is available and the techniques to be implemented are relevant;

- *The ability of the KDD process to learn as new data is received.* At the time being, when new data is added to the DW, the data mining techniques will run again and the discovered knowledge would be affected by deleting or modifying existing rules, or by adding new rules. Future research could be directed towards automating this process which makes the knowledge base able to learn and adapt as new data is received;

- *Comparative research studies.* Carry out the study using the ARDSSQ in different countries and compare the results. It is important to realize that some modifications need to be made to the ARDSSQ to reflect any country's specific educational system environment;

- *Different Application domains.* Research could take place using the new DSS methodology that is proposed by this research, however applied to different application domains. Examples are the Banking, Stock exchanges, and Market data.

References

Aas, K., Husbey, R., and Thune, M. (1999), ***Data Mining: A Survey***, unpublished report, Norwegian Computing Center, Norway.

Abiteboul, S., Hull, R., and Vianu, V. (1995), ***Foundations Of Databases***, Addison-Wesley Publishing Company, Reading, Massachusetts.

Abraham, T., and Wankel, C. (1995), 'Supporting Decision Support: Where Information On DSS Is Located', ***Decision Support Systems***, vol. 14, pp. 299-312.

Adam, F., Fahy, M., and Murphy, C. (1997), 'A Framework For The Classification Of DSS Usage Across Organisations', ***Decision Support Systems***, vol. 22, pp. 1-13.

Adamson, C., and Venerable, M. (1998), ***Data Warehouse Design Solutions***, John Wiley and Sons, Inc., New York.

Admission and Registration Department. (1995), ***Statistics From The Admission Records***, Arab Academy for Science and Technology & Maritime Transport (AASTMT), Alexandria.

Admission and Registration Department. (1997), ***Admission And Registration Statistics from 1990-1997***, Arab Academy for Science and Technology & Maritime Transport (AASTMT), Alexandria.

Adriaans, P., and Zantinge, D. (1996), ***Data Mining***, Addison Wesley Longman Limited, Harlow.

Agrawal, R., and Srikant, R. (2000), 'Privacy-Preserving Data Mining', In: ***Proceedings Of ACM SIGMOD '2000 International Conference On Management Of Data***, May 16-18, Dallas, Texas, pp. 439-450.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. (1996), 'Fast Discovery Of Association Rules', In Fayyad, U., Piatetsky, G., and Smyth, P. (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 307-328.

Agrawal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, and Sarawagi, S. (1996), 'On the Computation of Multidimensional Aggregates', In: ***Proceedings of the 22nd VLDB Conference***, Mumbay.

Ahn, J., and Ezawa, K., J. (1997), 'Decision Support For Real-Time Telemarketing Operation Through Bayesian Network Learning', ***Decision Support Systems***, vol. 21, pp. 17-27.

Akkus, A., and Güvenir, H. (1996), 'K Nearest Neighbor Classification On Feature Projections', In: ***Proceedings of the ICML' 96 Conference***, Bari, pp. 12-19.

Alavi, M., and Leidner, D. E. (2001), 'Knowledge Management And Knowledge Management Systems: Conceptual Foundations And Research Issues', ***MIS Quarterly***, vol. 25, no 1, pp. 107-136.

Ali, F., and Wallace, W. (1997), 'Bridging The Gap Between Business Objectives And Parameters Of Data Mining Algorithms', ***Decision Support Systems***, vol. 21, pp. 3-15.

Alter, S. (1977), 'A Taxonomy of Decision Support Systems', ***Sloan Management Review***, Fall, pp. 39-56.

Alter, S. (1992), ***Information Systems: A Management Perspective***, Addison Wesley Publishing Company, Reading, MA.

Anand, S., Bell, D., and Hughes, J. (1995), 'Evidence Based Discovery Of Knowledge In Databases', In: ***Colloquium on Knowledge Discovery in Databases***, The Institute of Electrical Engineers (IEE), London, pp. 9/1-9/4.

Anderberg, M. (1973), ***Cluster Analysis For Applications***, Academic Press Inc., London.

Anderson, D., Sweeney, D., and Williams, T. (1996), ***Statistics For Business And Economics***, West Publishing Company, St. Paul, MN.

Applegate, L. (1989), 'Executive Information Systems: Technology Overview', **Harvard *Business Review***, August-September, pp. 159-189.

Armstrong, D. (1992), 'The People Factor In EIS Success', In Watson, H., Rainer, R., and, Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 287-297.

Armstrong, G., and Kotler, P. (2000), ***Marketing: An Introduction***, 5th ed., Prentice Hall, New Jersey.

Baik, J. (2000), ***The Effect of CASE Tools on Software Development Effort***, Center for Software Engineering, Computer science Department, University of Southern California, California, unpublished.

Bandemer, H., and Gottwald, S. (1995), ***Fuzzy Sets, Fuzzy Logic, Fuzzy Methods with Applications***, John Wiley and Sons, Inc., West Sussex.

Barquin, R. (1997), 'A Data Warehousing Manifesto', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 3-16.

Barquin, R., Paller, A., and Edelstein, H. (1997), 'Ten Mistakes To Avoid For Data Warehousing Managers', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 145-

156.

Barr, S. and Sharda, R. (1997), 'Effectiveness Of Decision Support Systems: Development Or Reliance Effect', ***Decision Support Systems***, vol. 21, pp. 133-146.

Barron, T., and Saharia, A. (1995), 'Data Requirements In Statistical Decision Support Systems: Formulation And Some Results In Choosing Summaries', ***Decision Support Systems***, vol. 15, pp. 375-388.

Barrow, C. (1992), 'Implementing an Executive Information System: Seven Steps For Success', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 107-116.

Bee, R., and Bee, F. (1990), ***Management Information Systems And Statistics***, Institute of Personnel Management, London.

Beeri, C., Elber, G., Milo, T., Sagiv, Y., Shmueli, O., Tishby, N., Kogan, Y., Konopnicki, D., Mogilevski, P., and Slonim, N. (1998), 'Web Suite- A Tool Suite For Harnessing Web Data', In: ***Proceedings Of The International Workshop On The Web and Databases***, Valencia.

Benjamin, R., and Blunt, J. (1999), 'Critical Information Technology Issues In The Year 2000', In Galliers, R., Leidner, D., and Bernadette S. Baker (Eds), ***Strategic Information Management: Challenges And Strategies In Managing Information Systems***, Butterworth Heinemann, Oxford, pp. 161-186.

Bennett, J. (1983), ***Building Decision Support Systems***, Addison-Wesley, Reading, Massachusetts.

Berson, A. (1996), ***Client/Server Architecture***, McGraw-Hill, New York.

Berson, A., and Smith, S. (1997), ***Data Warehousing, Data Mining, and OLAP***, McGraw-Hill, New York.

Bhargava, H., Krishnan, R., and Muller, R. (1997), 'Decision Support on Demand: Emerging Electronic Markets For Decision Technologies', ***Decision Support Systems***, vol. 19, pp. 193-214.

Blockeel, H., and De Raedt, L. (1997), ***Top-Down Induction of Logical Decision Trees***, Technical Report CW 247, Katholieke Universiteit Leuven, Department of Computer Science, K.U. Leuven, Heverlee.

Boisot, M. (1998), ***Knowledge Assets: Securing Competitive Advantage In The Information Economy***, Oxford.

Boudreau, M., Gefen, D., and Straub, D. W. (2001), 'Validation in Information Systems Research: A State-of-the-Art Assessment', ***MIS Quarterly***, vol. 25, no 1, pp. 1-16.

Brace, N., Kemp R., and Snelgar, R. (2000), ***SPSS For Psychologists A Guide To Data Analysis Using SPSS For Windows***, Lawrence Erlbaum Associates Inc. Publishers, New Jersey.

Brachman, R., and Anand, T. (1996), 'The Process Of Knowledge Discovery In Databases', In Fayyad, U., Piatetsky, G., and Padharic Smyth (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 37-57.

Brislin, R. (1970), 'Back translation for cross-cultural research', ***Journal of Cross-Cultural Psychology***, vol. 1, pp. 185-216.

Brislin, R. (1976), ***Translation: Application and Research***, John Wiley, New York.

Broadbent, M., Weill, P., and Neo, A. (1999), 'Strategic Context And Patterns Of IT Infrastructure Capability', ***Journal of Strategic Information Systems***, vol. 8, pp. 157-187.

Bryman, A., and Cramer, D. (1999), ***Quantitative Data Analysis With SPSS Release 8 For Windows***, Routledge, London.

Buntine, W. (1996), 'Graphical Models For Discovering Knowledge', In Fayyad, U., Piatetsky, G., and Padharic Smyth (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 59-81.

Burn-Throronton, K., Garibaldi, J., and Hamer, P. (1998), 'Data Mining Algorithms: Applicability To Data Sets', In: ***Proceedings Of The INC 98 Conference***, Plymouth.

Burn-Throronton, K., Thorpe, S., and Edenbrandt, L. (1999), 'Improving Clinical Decision Support Systems Using Data Mining', In: ***Proceedings of The SPIE Conference***.

Canavos, G., and Miller, D. (1995), ***Modern Business Statistics***, Duxbury.

Carmichael, J., W., and Sneath, P. H. A. (1969), 'Taxometric Maps', ***Syst. Zool***, vol. 18, pp. 402-415.

Cavaye, A. (1995), 'User Participation In System Development Revisited', ***Information & Management***, vol. 28, pp. 311-323.

Celko, J. (1995), ***Instant SQL Programming***, WROX Press Ltd, Birmingham.

Challenger, J., Iyengar, A., and Dantzig, P. (1999), 'A Scalable System for Consistency Caching Dynamic Web Data', In: ***proceedings of the IEEE INFOCOM '99***

Conference.

Chan, Y., Huff, S., and Copeland, D. G. (1998), 'Assessing Realized Information Systems Strategy', ***Journal of Strategic Information Systems***, vol. 6, pp. 273-298.

Channon, D. (1998), 'The Strategic Impact Of IT On The Retail Financial Services Industry', ***Journal of Strategic Information Systems***, vol. 7, pp. 183-197.

Chaston, I. (2001), **E-Marketing Strategy**, McGraw-Hill, Berkshire.

Chen, C., and Paul, R., J. (2001), 'Visualizing A Knowledge Domain's Intellectual Structure', ***IEEE Computer***, vol. 34, no 3, pp. 65-71.

Chen, H., and Sinha, D. (1996), 'An Inventory Decision Support System Using The Object-Oriented Approach', ***Computers Operations Research***, vol. 23, no 2, pp. 153-170.

Chen, M., Han, J., and Yu, P. S. (1996), 'Data Mining: An Overview From A Database Perspective', ***IEEE Knowledge And Data Engineering***, vol. 8, no 6, pp. 866-883.

Cheng, A. (1995), 'The UK Stock Market And Economic Factors: A New Approach', ***Journal Of Business Finance And Accounting***, vol. 22, no 1, pp. 129-143.

Chengular-Smith, I., Ballou, D., and Pazer, H. L. (1999), 'The Impact Of Data Quality Information On Decision Making: An Exploratory Analysis', ***IEEE Transactions On Knowledge And Data Engineering***, vol. 11, no 6, pp. 853-864.

Child, D. (1995), ***The Essentials Of Factor Analysis***, 2nd ed., Cassell Educational Limited, London.

Churchill, G. (1979), 'A Paradigm For Developing Better Measures Of Marketing Constructs', ***Journal of Marketing Research***, vol. 16, pp. 64-73.

Cliff, N. (1987), ***Analyzing Multivariate Data***, Harcourt Brace Jovanovich, Inc., Florida.

Clifton, C., and Marks, D. (1996), 'Security And Privacy Implications Of Data Mining', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec.

Codd, E. (1990), ***The Relational Model For Database Management Version 2***, Addison-Wesley Publishing Company, Reading, Massachusetts.

Cool: Gen 5.0 CASE Tools Manuals. (1997), different volumes, Sterling Software, London.

Cooper, B., Watson, H., Wixom, B., and Goodhue, D. L. (2000), 'Data Warehousing

Supports Corporate Strategy At First American Corporation', ***MIS Quarterly***, vol. 24, no 4, pp. 547-567.

Cooper, D., and Schindler, P. S. (1998), ***Business Research Methods***, Irwin/McGraw-Hill, Singapore.

Cormen, T., Leiserson, C., and Rivest, R. L. (2000), ***Introduction To Algorithms***, 24th printing, MIT Press, Massachusetts.

Corr, B. (1995), ***Essential Elements Of Business Information Systems***, The Guernsey Press Co. Ltd, Vale, Guernsey.

Coyle, J., Bardi, E., and Langley, J. (1992), ***The Management of Business Logistics***, 5th ed., West Publishing Company, St. Paul, MN.

Crispin, L. (2001), 'Extreme Rules Of The Road', ***The STQE Magazine***, vol. 3, no 4, pp. 24-30.

Croft, A., and Davison, R. (1999), ***Mathematics For Engineers A Modern Interactive Approach,*** Addison-Wesley Longman Ltd., Essex.

Czerniawska, F., and Potter, G. (1998), ***Business In A Virtual World: Exploring Information For Competitive Advantage***, Macmillan, London.

Daniel, A., and Kriegel, H. (1996), 'Visualization Techniques For Mining Large Databases: A Comparison', ***IEEE Knowledge And Data Engineering***, vol. 8, no 6, pp. 923-938.

Data Warehouse Administration. (1999), SAS In Association With The Data Warehouse Network [online], URL: http://**www.sas.com**

Database Models. (2001), David R. Frick & Company [online], URL: http://**www.frick-cpa.com/ss7/theory_models.asp**

Date, C. (1995), ***An Introduction To Database Systems***, 6th ed., Addison-Wesley Publishing Company, Reading, Massachusetts.

Dauphinais, G. (1987), 'The Information Draught In The Executive Suite', ***Price Waterhouse Review***, vol. 1, pp. 41-47.

David, W., Vitcent, T., Ada, W., and Fu, Y. (1996), 'Efficient Mining Of Association Rules In Distributed Database', ***IEEE Knowledge And Data Engineering***, vol. 8, no 6, pp. 911-922.

Davids, A. (1992), ***Practical Information Engineering: The Management Challenge***, Pitman Publishing, London.

Delis, A., and Roussopoulos, N. (1992), 'Performance And Scalability Of Client-Server Database Architecture', In: ***Proceedings Of The 18th VLDB Conference***, British Columbia University, Vancouver.

DeVellis, R. (1991), ***Scale Development: Theory And Applications***, SAGE Publications, Inc., London.

Devlin, B. (1997), ***Data Warehouse: From Architecture To Implementation***, Addison-Wesley Longman, Inc., Reading, Massachusetts.

Dewire, D. (1998), ***Thin Clients***, McGraw-Hill.

Dowling, C., E., and Kockemeyer, C. (2001), 'Automata For The Assessment Of Knowledge', ***IEEE Transactions On Knowledge And Data Engineering***, vol. 13, no 3, pp. 451-461.

Dzeroski, S. (1996), 'Inductive Logic Programming And Knowledge Discovery In Database', In Fayyad, U., Piatetsky, G., and Padharic Smyth (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 117-152.

Eades, D., C. (1965), 'The Inappropriateness Of The Correlation Coefficient As A Measure Of Taxonomic Resemblance', ***Syst. Zool.***, vol. 14, pp. 98-100.

Edelstein, H. (1997), 'An Introduction To Data Warehousing', In Barquin, R., and Herb Edlestein (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 31-50.

Edwards, J. (1999), ***3-Tier Client/Server At Work***, John Wiley and Sons, Inc., New York.

Elam, J., Jarvenpaa, S., and Schkade, D. A. (1992), 'Behavioral Decision Theory And DSS: New Opportunities for Collaborative Research', In Stor, E., and Benn R. Konsynski (Eds), ***Information Systems and Decision Making***, IEEE, pp. 51-74.

El-Kot, G. (2001), ***Team Player Styles, Team Design Variables, and Team Work Effectiveness In Egypt***, unpublished PhD Thesis, University of Plymouth Business School, Plymouth.

Elmasri, R., and Navathe, S. B. (2000), ***Fundamentals of Database Systems***, 3rd ed., Addison Wesley, Massachusetts.

El-Sawy, O. (1985), 'Personal Information Systems For Strategic Scanning In Turbulent Environment: Can The CEO Go On-Line?', ***MIS Quarterly***, vol. 1, no 1, pp. 53-60.

Everitt, R. (1980), ***Cluster Analysis***, 2nd ed., Halsted press- A division of John Wiley and Sons, Inc., London.

Fady, R. (2000), ***Electronic Payment Student Registration Systems***, unpublished MBA Thesis, The Advanced Management Institute (AMI), Alexandria.

Farbey, B., Land, F., and Targett, D. (1999), 'Moving IS Evaluation Forward: Learning Themes And Research Issues', ***Journal Of Strategic Information Systems***, vol. 8, pp. 189-207.

Farkas, C., and Wetlaufer, S. (1996), 'The Ways Chief Executive Officers Lead', ***Harvard Business Review***, March-April, pp. 110-122.

Fayyad, U., Piatetsky, G., and Smyth, P. (1996), 'From Data Mining To Knowledge Discovery: An Overview', In Fayyad, U., Piatetsky, G., and Smyth, P. (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 1-34.

Fink, A. (1995a), ***How To Ask Survey Questions***, Thousand Oaks, Sage, California.

Fink, A. (1995b), ***The Survey Handbook***, Thousand Oaks, Sage, California.

Firestone, J. (1998a), ***Dimensional Modeling And E-R Modeling In The Data Warehouse***, EIS Inc.

Firestone, J. (1998b), ***Dimensional Object Modeling***, EIS Inc.

Fitzgerald, E. (1993), 'Success Measures For Information Systems Strategic Planning', ***Journal of Strategic Information Systems***, vol. 2, pp. 335-349.

Foster, I., Insley, J., Laszewski, G., Kesselman, C., and Thiebaux, M. (1999), 'Distance Visualization: Data Exploration On The Grid', ***IEEE Computer***, vol. 32, no 12, pp. 36-43.

Frasconi, P., Gori, M, and Soda, G. (1999), 'Data Categorization Using Decision Trellises', ***IEEE Transactions On Knowledge And Data Engineering***, vol. 11, no 5, pp. 697-712.

Friend, D. (1992), 'EIS And The Collapse Of The Information Pyramid', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***. John Wiley and Sons, Inc., New York, pp. 327-335.

Fukuda, T., Morimoto, Y., and Morishita, S. (1996), 'Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec, pp. 13-23.

Funkhouser, T. (1995), 'RING: A Client-Server System For Multi-User Virtual Environment', In: ***Proceedings Of The ACM SIGGRAPH Conference***, New York, pp. 85-92.

Gaines, B. (1996), 'Transforming Rules And Trees', In Fayyad, U., Piatetsky, G., and

Smyth, P. (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 205-220.

Galbraith, J. (1974), 'Organization Design: An Information Processing View', ***Interfaces***, pp. 28-36.

Games, P., and Lucas, P. (1966), 'Power Of The Analysis Of Variance Of Independent Groups On Non-Normal and Normally Transformed Data', ***Educational and Psychological Measurement***, vol. 26, pp. 311-327.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999), 'Mining Very Large Databases', ***IEEE Computer***, vol. 32, no 8, pp. 38-45.

Garcia-Molina, H., Labio, W., and Yang, J. (1998), 'Expiring Data In A Warehouse', In: ***Proceedings Of The 24th VLDB Conference***, New York.

Geiger, J. (1992), 'Management Implications Of Installing A Modern Financial Aid System', ***Journal of Systems Management***, vol. 43, no 3, pp. 6-9.

Ghiselli, E., Campbell, J., and Zedeck, S. (1981), ***Measurement Theory for the Behavioral Sciences***, W. H. Freeman, San Francisco.

Global Education and Training Information Service-Egyptian Universities. (1999), The British Council, London.

Gordon, A. D. (1981), ***Monographs On Applied Probability And Statistics Classification***, The University Press, Cambridge.

Gower, J. C. (1966), 'Some Distance Properties Of Latent Root And Vector Methods Used In Multivariate Analysis', ***Biometrics***, vol. 53, no 4, pp. 325-338.

Gower, J. C. (1971), 'A General Coefficient Of Similarity And Some Of Its Properties', ***Biometrics***, vol. 27, no 4, pp. 857-874.

Gray, P, Alter, S., DeSanctis, G., Dickson, G., Johansen, R., Kraemer, K., Olfman, L. and Vogel, D. R. (1992), 'Group Decision Support Systems', In Stor, E., and Konsynski, B., R. (Eds), ***Information Systems And Decision Making***, IEEE, pp. 75-136.

Gray, P., and Watson, H., J. (1999), ***The New DSS: Data warehouse, OLAP, MDD, and KDD***, Unpublished report, Georgia State University, Georgia.

Green, P., E., Donalds, T., and Albaum, G. (1988), **Research For Marketing Decisions**, 5th ed., Prentice Hall, New Jersey.

Guha, S., Rastorgi, R., and Shim, K. (2000), 'ROCK: A Robust Clustering Algorithm For Categorical Attributes', ***Information Systems***, vol. 25, no 5, pp. 345-366.

Gulden, G., and Ewers, D. (1992), 'Is your ESS meeting the need?', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems:***

Emergence, Development, Impact, John Willey and Sons, Inc., New York, pp. 117-125.

Gupta, U., G. (1996), **Managing Information Systems: A Managerial Perspective, West Publishing Company**, St. Paul, MN.

Guyon, I., Matic, N., and Vapnik, V. (1996), 'Discovering Informative Patterns And Data Cleaning', In Fayyad, U., Piatetsky, G., and Smyth, P. (Eds), ***Advances In Knowledge Discovery And Data Mining***, AAAI Press/ The MIT Press, California, pp. 181-203.

Hackett, G. and Luffrum, P. (1999), ***Business Decision Analysis***, Blackwell Publishers, Oxford.

Hadden, E. (1998a), ***Building For Successful Data Warehouses And Data Marts***, BBS International, Cairo.

Hadden, E. (1998b), ***Planning For Successful Data Warehouses And Data Marts***, BBS International, Cairo.

Hair, J., Anderson, R., Tatham, R., and Black, W. (1995), ***Multivariate Data Analysis With Readings***, 4th ed., Prentice-Hall International, Inc., New Jersey.

Han, J. (1996), 'Data Mining Techniques', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec, pp. 545-545.

Han, J., Fu, Y., Wang, W., Krzysztof, K., and Zaiane, O. (1996), 'DMQL: A Data Mining Query Language For Relational Databases', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec.

Han, J., Lakshmanan, L., and Ng, R., T. (1999), 'Constraint-Based, Multidimensional Data Mining', ***IEEE Computer***, vol. 32, no 8, pp. 46-50.

Harinarayan, V., Rajaraman, A., and Ullman, J. (1996), 'Implementing Data Cubes Efficiently', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec, pp. 205-216.

Harrington, J. (1998), ***SQL Clearly Explained***, AP Professional, Massachusetts.

Hartigan, J. (1975), ***Clustering Algorithms***, John Wiley and Sons, Inc., New York.

Hawk, S., and Bariff, M., L. (1995), 'An Examination Of Organizational Strategies For Supporting DSS', ***Information & Management***, vol. 28, pp. 77-88.

Hawkins, C., and J., Weber. (1980), **Statistical Analysis: Applications To Business and Economics**, Harper and Row Publishers, New York.

Hicks, J. (1993), ***Management Information Systems: A User Perspective***, 3rd ed., West Publishing Company, St. Paul, Minneapolis.

Hogue, J., and Hugh Watson. (1983), 'Management's Role In The Approval And Administration Of Decision Support Systems', ***MIS Quarterly***, vol. 3, pp. 15-26.

Holsapple, C. W., and Whinston, A., B. (1996), ***Decision Support Systems A Knowledge-Based Approach***, West Publishing Company, St. Paul, Minneapolis.

Houdeshel, G., and Watson, H. (1992), 'The Management Information Decision Support-MIDS- System At Lockheed-Georgia', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 13-31.

Hout, T., and Carter, J. (1995), 'Getting It Done: New Roles For Senior Executives', ***Harvard Business Review***, June-July, pp. 133-145.

Hsu, Y. (1999), ***Recruitment and Selection and Human Resource Management in the Taiwanese Cultural Context***, unpublished PhD Thesis, University of Plymouth Business School, Plymouth.

Huber, G. (1999), 'The Effect Of Advanced Information Technologies On Organisational Design, Intelligence, And Decision Making', In Galliers, R., Leidner, D., and Baker, B., S. (Eds), ***Strategic Information Management: Challenges And Strategies In Managing Information Systems***, Butterworth-Heinemann, Oxford, pp. 487-522.

Hufford, D. (1997), 'Metadata Repositories: The Key To Unlocking Information In Data Warehouses', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 225-262.

Hui, C. H. and Triandis, H. C. (1985), 'Measurement in cross-cultural psychology. A review and comparison of strategies', ***Journal of Cross-Cultural Psychology***, vol. 16 (2), pp. 131-152.

Humphries, M., Hawkins, M., and Dy, M., C. (1999), ***Data Warehousing Architecture And Implementation***, Prentice Hall PTR, New Jersey.

Hunt, E. B., Marin, J., and Stone, P., J. (1966), ***Experiments In Induction***, Academic Press, New York.

Imielinski, T. (1996), 'From File Mining To Database Mining', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec.

Inmon, W. (1993), ***Building the Data Warehouse***, QED publishing Group.

Inmon, W., and Hackathorn, R., D. (1994), ***Using The Data Warehouse***, John Wiley and Sons, Inc., New York.

Inmon, W., Rudin, K., Buss, C., and Sousa, R. (1999), ***Data Warehouse Performance***, John Wiley and Sons, Inc., Toronto.

Jaeger, M., Mannila, H., and Weydert, E. (1996), 'Data Mining As Selective Theory Extraction In Probabilistic Logic', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec.

Jain, A. and Dubes, R., C. (1988), ***Algorithms for Clustering Data***, Prentice Hall, New Jersey.

Jajodia, S., and Sandhu, R. (1991), 'A Novel Decomposition Of Multilevel Relations Into Single-Level Relations', In: ***Proceedings Of IEEE Symposium On Security And Privacy, Oakland California, May***, pp. 300-313.

Jardine, N., and Sibson, R. (1968), 'The Construction Of Hierarchic And Non-Hierarchic Classifications', ***Comp. J.***, vol. 11, pp. 117-184.

Johnson, C., Parker, S., Hansen, C., Kindlmann, G., and Livnat, Y. (1999), 'Interactive Simulation And Visualization', ***IEEE Computer***, vol. 32, no 12, pp. 59-65.

Jordan, D., and Smith, P. (1997), ***Mathematical Techniques***, 2nd ed., Oxford.

Kanji, G. (1999), ***100 Statistical Tests***, New Edition, SAGE Publications Ltd., London.

Keil, M., Mann, J., and Rai, A. (2000), 'Why Software Projects Escalate: An Empirical Analysis And Test of Four Theoretical Models', ***MIS Quarterly***, vol. 24, no 4, pp. 631-664.

Keyes, J. (1993), ***Infotrends: The Competitive Use Of Information***, McGraw-Hill, New York.

Khosla, I., Kuhn, B., and Soparkar, N. (1996), 'Database Search Using Information Mining', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec.

Kimball, R. (1996), ***The Data Warehouse Toolkit: Practical Techniques For Building Dimensional Data Warehouses***, John Wiley and Sons, Inc., Toronto.

Klein, M., and Methlie, L., B. (1995), ***Knowledge-Based Decision Support Systems With Applications In Business***, 2nd ed., John Wiley and Sons Ltd., West Sussex.

Kline, P. (1993), ***Personality: The Psychometric View***, Routledge, London.

Kline, P. (1998), ***The New Psychometrics Science, Psychology, And Measurement***,

Routledge, London.

Kononenko, I., Bratko, I., and Roskar, E. (1984), ***Experiments In Automatic Learning Of Medical Diagnosis Rules***, Josef Stefan Institute, Ljubljana.

Konsynski, B., Stohr, E., and McGee, J., V. (1992), 'Review And Critique Of DSS', In Stor, E., and Konsynski, B., R., ***Information Systems And Decision Making***, IEEE, pp. 7-26.

Kumar, A. (2000), ***Global Executive Information Systems: Key Issues And Trends***, Garland Publishing, Inc., New York.

Lance, G., N., and Williams, W. T. (1966), 'Computer Programs for Hierarchical Polythetic Classification ('Similarity Analyses')', ***Computer J.***, vol. 9, pp. 60-64.

Larsen, B., and Aone, C. (2000), 'Fast And Effective Text Mining Using Linear-time Document Clustering', In: ***Proceedings Of ACM SIGMOD '2000 International Conference On Management Of Data***, May 16-18, Dallas, Texas, pp. 16-22.

Laudon, K., and Laudon, J., P. (1998), ***Management Information Systems: New Approach to Organization and Technology***, 5th ed., Prentice Hall International, Inc., New Jersey.

Laudon, K., and Laudon, J., P. (2000), ***Business Information Systems: A Problem-Solving Approach***, The Dryden Press, Chicago.

Laudon, K., and Laudon, J., P. (2001), ***Essentials of Management Information Systems: Organization and Technology in the Networked Enterprise***, 4th ed., Prentice Hall International, Inc., New Jersey.

Lawrence, P. and Dayer, D. (1983), ***Renewing American Industry***, The Free Press, New York.

Lawrence, P. and Lorsch, J. (1967), ***Organization And Environment: Managing Differentiation And Integration***, Harvard Business School Press, Boston.

Lederer, A., and Sethi, V. (1988), 'The Implementation Of Strategic Information Systems Planning Methodologies', ***MIS Quarterly***, vol. 12, no 3, pp. 445-461.

Lee, A., and Rundensteiner, E., A. (1998), 'Data Warehouse Evolution: Consistent Metadata Management', In: ***Proceedings Of The IEEE Conference on SMC***, San Diego, California.

Lee, D., Newman, P., and Price, R. (1999), ***Decision Making In Organisations***, Financial Times Management, London.

Lee, H. and Clark, T., H. (1999), 'Strategies In Response To The Potential Of

Electronic Commerce', In Galliers, R., Leidner, D., and Baker, B., S. (Eds), ***Strategic Information Management: Challenges And Strategies In Managing Information Systems***, Butterworth-Heinemann, Oxford, pp. 397-425.

Leidner, D. (1999), 'Understanding Information Culture: Integrating Knowledge Management Systems Into Organisations', In Galliers, R., Leidner, D., and Baker, B., S. (Eds), ***Strategic Information Management: Challenges And Strategies In Managing Information Systems***, Butterworth-Heinemann, Oxford, pp. 523-550.

Leidner, D., and Fuller, M. (1997), 'Improving Student Learning Of Conceptual Information: GSS Supported Collaborative Learning Vs. Individual Constructive Learning', ***Decision Support Systems***, vol. 20, pp. 149-163.

Levin, E. (1997), 'Developing A Data Warehousing Strategy', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 53-89.

Levinson, H. (1996), 'When Executives Burn Out', ***Harvard Business Review***, July-August, pp. 152-163.

Lim, J., and O'Connor, M. (1996), 'Judgmental Forecasting With Interactive Forecasting Support Systems', ***Decision Support Systems***, vol. 16, pp. 339-357.

Lin, T., and Pourahmadi, M. (1998), 'Non-Parametric And Non-Linear Models And Data Mining In Time Series: A Case Study On The Canadian Lynx Data', ***Journal of The Royal Statistical Society SERIES C***, vol. 47, no 2, pp. 187-201.

Liu, B., Hsu, W., and Ma, Y. (2000), 'Pruning And Summarizing The Discovered Associations', In: ***Proceedings Of ACM SIGMOD '2000 International Conference On Management Of Data***, May 16-18, Dallas, Texas, pp. 125-134.

Liu, B., Hsu, W., Mun, L., and Lee, H., Y. (1999), 'Finding Interesting Patterns Using User Expectations', ***IEEE Transactions On Knowledge And Data Engineering***, vol. 11, no 6, pp. 817-832.

Livingston, G., and Rumsby, B. (1997), 'Database Design For The Data Warehouses: The Basic Requirements', In Barquin, R., and Edlestein, H. (Eds), ***Planning and Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 179-198.

London Internet Exchange (LINX). (2001), [online], URL:

http://www.linx.net/press/newsletter/issue4-summer/volume4.pdf

Long, L. (1989), ***Management Information Systems***, Prentice Hall, New Jersey.

Long, L., and Long, N. (2001), ***Computers***, 8th ed., Prentice Hall, Upper Saddle River, New Jersey.

Lord, F. M. (1953), 'On The Statistical Treatment Of Football Numbers', ***American Psychologist***, vol. 8, pp. 750-751.

Makram, H. (2000), ***Data Warehousing For Student Recruitment Systems***, unpublished MBA Thesis, RITI, Cairo.

Marakas, G. (1998), ***Decision Support Systems In The 21st Century***, 1st ed., Prentice Hall, New Jersey.

Martella, R., Nelson, R., and Marchand-Martella, N. (1999), ***Research Methods***, Allyn and Bacon, Boston, Massachusetts.

Martin, E., Dehayes, D., W., Hoffer, J., A., and Perkins, W., C. (1994), ***Managing Information Technology***, Macmillan.

Mason, R., Lind, D., and Marchal, W. (1999), ***Statistical Techniques In Business And Economics***, 10th ed., Irwin/McGraw-Hill, Boston, Massachusetts.

Mattison, R. (1996), ***Data warehousing Strategies, Technologies, And Techniques***, McGraw-Hill, Indiana.

Mattison, R. (1997), ***Data Warehousing And Data Mining For Telecommunications***, Artech House, Inc., Norwood, Massachusetts.

MCDBA SQL Server 7.0 Administration Study Guide. (1999), Osborne McGraw-Hill, California.

McFarland, G., Rudmik, A., and Operandi, M. (1999), ***Object-Oriented Database Management Systems Revisited***, DACS, Florida.

McGehee, B., Miller, C., Shepker, M, and Bersinic, D. (1998), ***MCSE SQL Server 6.5 Administration***, Sams Publishing, Indianapolis, IN.

McGregor, R. (1999), ***Using C*** ++, QUE, Indiana.

McLeod, R., and Jones, J. (1992), 'Making Executive Information Systems More Effective', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 53-69.

McNurlin, B. (1987), 'Executive Information Systems', ***EDP ANALYZER***, vol. 25, no 4, pp. 1-11.

McSherry, D. (1997), 'Knowledge Discovery By Inspection', ***Decision Support Systems***, vol. 21, pp. 43-47.

Mi, P., and Scacchi, W. (1996), 'A Meta-Model For Formulating Knowledge-Based Models Of Software Development', ***Decision Support Systems***, vol. 17, pp. 313-330.

Michie, D. (1986), ***On Machine Intelligence***, 2nd ed., Ellis Horwood, Chichester.

Millet, I., and Mawhinney, C. (1992), 'Executive Information Systems: A Critical Perspective', ***Information & Management***, vol. 23, pp. 83-92.

Millet, I., Mawhinney, C., and Kallman, E. (1992), 'A Path Framework For Executive Information System', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***. John Wiley and Sons, Inc., New York, pp. 127-144.

Milne, R., and Nelson, C. (1995), 'Knowledge Guided Data Mining', In: ***Colloquium On Knowledge Discovery In Databases***, The Institute of Electrical Engineers (IEE), London, pp. 6/1-6/3.

Mimno, P. (1997), 'Data Warehousing Architectures', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 159-177.

Mimno, P. (1999), 'Build your Data Warehouse Right The First Time', Brio Technology [online], URL: http://**www.brio.com**

Mintzberg, H. (1973), ***The Nature Of Managerial Work***, Harper and Row, New York.

Mintzberg, H. (1981), 'What Is Planning Anyway?', ***Strategic Management Journal***, vol. 2, pp. 319-324.

Mintzberg, H. (1992), 'The Manager's Job: Folklore And Facts', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 35-52.

Morgan, B. (1981), 'Three Applications Of Methods Of Cluster-Analysis', ***The Statistician***, vol. 30, no 3, pp. 205-223.

Moulet, M., and Kodratoff, Y. (1995), 'From Machine Learning Towards Knowledge Discovery In Databases', In: ***Colloquium On Knowledge Discovery In Databases***, The Institute of Electrical Engineers (IEE), London, pp. 5/1-5/3.

Munton, A., Silvester, J., Stratton, P., and Hanks, H. (1999), ***A Practical Approach to Coding Qualitative Data***, John Wiley and Sons, Inc., New York.

Neal, D. (1997), 'How to Justify the Data Warehouse and Gain the Top Management Support', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 91-115.

Newmark, P. (1988), ***Approaches to Translation***, 2nd Ed., Prentice Hall, New York.

Nunnally, J. (1978), ***Psychometric Theory***, McGraw-Hill, New York.

O'brien, J. (1996), ***Introduction To Information Systems***, Irwin, Chicago.

O'Driscoll, T., Massey, A., and Montoya-Weiss, M., M. (1999), 'Virtual Mentor: Enabling Knowledge Management Through An Electronic Performance Support System', SIM International paper award competition-2nd place [online], URL: **http://www.simnet.org**

Onder, J., and Nash, T. (1999), 'The Approach To Building A Business Data Warehouse', SYSIX [online], URL: **http://www.sysix.com**

Orfali, R., Karkey, D., and Edwards, J. (1999), ***Client/Server Survival Guide***, John Wiley and Sons, Inc., New York.

Ortuzar, J. D., and Willumsen, L., G. (1994), ***Modelling Transport***, John Wiley and Sons Ltd, New York.

Paller, A. (1997), 'A Roadmap To Data Warehousing', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 17-29.

Palvia, P., Kumar, A., Kumar, N., and Hendon, R. (1996), 'Information Requirements Of A Global EIS: An exploratory Macro Assessment', ***Decision Support Systems***, vol. 16, pp. 169-179.

Parrish, A., and Zweben, S. (1991), 'Analysis And Refinement Of Software Test Data Adequacy Properties', ***IEEE Transactions On Software Engineering***, vol. 17, no 6, pp. 565-581.

Parsaye, K., Chignell, M., Khoshafian, S., and Wong, H. (1989), ***Intelligent Databases: Object-Oriented, Deductive, Hypermedia Technologies***, John Wiley and Sons, Inc., New York.

Patterson, A., and Niblett, T. (1983), ***ACLS User Manual***, Intelligent Terminals Limited, Glasgow.

Pegels, C. C. (1995), ***Total Quality Management: A Survey Of Its Important Aspects***, Boyed and Fraser Publishing, Danvers, MA.

Power, D., J. (1999), 'Decision Support Systems Glossary', DSS Resources [online], URL: **http://dssresources.com/glossary/1999**

Pratt, P., and Adamski, J. (1987), ***Database Systems Management and Design***, Boyd and Fraser Publishing.

Probst, G., Raub, S., and Romhardt, K. (2000), ***Managing Knowledge Building Blocks for Success***, John Wiley and Sons, Inc., New York.

Pyle, D. (1999), ***Data Preparation For Data Mining***, Morgan Kaufmann Publishers, Inc., California.

Quinlan, J., R. (1986), 'Induction Of Decision Trees', ***Machine Learning***, vol. 1, pp. 81-106.

Quinlan, J., R. (1987), 'Simplifying Decision Trees', ***Int. J. Man-Machine Studies***, vol. 27, pp. 221-234.

Quinlan, J., R. (1993), ***C 4.5: Programs For Machine Learning***, Morgan Kaufmann Publishers, London.

Raden, N. (1997), 'Choosing The Right OLAP Technology', In Barquin, R., and Edlestein, H. (Eds), ***Planning And Designing The Data Warehouse***, Prentice Hall, New Jersey, pp. 199-224.

Raghunathan, S. (1996), 'A Structured Modeling Based Methodology To Design Decision Support Systems', ***Decision Support Systems***, vol. 17, pp. 299-312.

Ragowsky, A., Ahituv, N., and Neumann, S. (1996), 'Identify The Value And Importance Of An Information System Application', ***Information & Management***, vol. 31, pp. 89-102.

Ramakrishnan, N. and Grama, A. (1999), 'Data mining: From Serendipity To Science', ***IEEE Computer***, vol. 32, no 8, pp. 34-37.

Ramirez, R., Kulkarni, U., and Moser, K. (1996), 'Derived Data For Decision Support Systems', ***Decision Support Systems***, vol. 17, pp. 119-140.

Rasmussen, E. (1992), 'Clustering Algorithms', In Frakes, W., and Baeza-Yates, R. (Eds), ***Information Retrieval Data Structures & Algorithms***, Prentice-Hall, pp.419-442.

Ravichandran, T., and Rai, A. (2000), 'Quality Management In Systems Development: An Organizational System Perspective', ***MIS Quarterly***, vol. 24, no 3, pp. 381-415.

Rawlings, J., Pantula, S., and Dickey, D., A. (1998), ***Applied Regression Analysis A Research Tool***, 2nd ed., Springer-Velrlag, Inc., New York.

Regan, E., and O'cnnor, B., N. (1994), ***End-User Information System***, Macmillan.

Reynolds, G. (1995), ***Information Systems For Managers***, West Publishing company, St. Paul, MN.

Reynolds, N., Diamantopoulos, A., and Schlegelmilch, B. (1993), 'Pretesting In Questionnaire Design: A Review Of The Literature And Suggestions For Further Research', ***Journal Of The Market Research Society***, vol. 35, pp. 171-183.

Riedel, E., Faloutsos, C., Ganger, G., and Nagle, D. (2000), 'Data Mining On An OLTP System (Nearly) For Free', In: ***Proceedings Of ACM SIGMOD '2000 International Conference On Management Of Data***, May 16-18, Dallas, Texas, pp. 13-21.

Robson, W. (1997), ***Strategic Management & Information Systems***, 2nd ed., Financial Times Management-Pitman Publishing, London.

Rockart, J., and DeLong, D. (1988), ***Executive Support Systems: The Emergence Of Top Management Computer Use***, Dow Jones-Irwin, Homewood, IL.

Rockart, J., and Treacy, M. (1992), 'The CEO Goes On-Line', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 3-12.

Rosner, D., Grote, B., Hartman, K., and Kofling, B. (1998), 'From Natural Language Documents To Sharable Product Knowledge: A Knowledge Engineering Approach', In Borghoff, U. and Pareschi, R. (Eds), ***Information Technology for Knowledge Management***, Springer, pp. 151-182.

Rowley, J. (1996), ***The Basics Of Information Systems***, Library Association Publishing.

Rust, J. and Golombok, S. (1999), ***Modern Psychometrics***, 2nd ed., Routledge, London.

Sadler, A. J., and Thorning, D., W., S. (1999), ***Understanding Pure Mathematics***, Oxford University Press, Oxford.

Sanders, G., and Courtney, J., F. (1985), 'A Field Study of Organisational Factors Influencing DSS Success', ***MIS Quarterly***, vol. 2, pp. 77-92.

Saraee, M., and Theodoulidis, B. (1995), 'Knowledge Discovery In Temporal Databases', In: ***Colloquium On Knowledge Discovery In Databases***, The Institute of Electrical Engineers (IEE), London, pp. 1/1-1/4.

Saunders, M., Lewis, P., and Thornhill, A. (1997), ***Research Methods For Business Students***, Financial Times-Pitman Publishing, London.

Schildt, H. (1998), ***The Complete Reference C*** ++, 3rd ed., Osborne McGraw-Hill, California.

Schneider, G., Bruell, S. (1992), ***Concepts In Data Structures And Software Development***, West Publishing Company, St. Paul, MN.

Sedgewick, R. (1998), ***Algorithms in C*** ++, 3rd ed., Addison-Wesley.

Selfridge, P., Srivastava, D., and Wilson, L. (1996), 'IDEA: Interactive Data Exploration And Analysis', In: ***Proceedings Of ACM SIGMOD '96***

International Conference On Management Of Data, June 4-6, Montreal, Quebec, pp. 24-34.

Senn, J. (1978), ***Information Systems In Management***, Wadsworth Publishing Company.

Senn, J. (1989), ***Analysis And Design Of Information Systems***, 2nd ed., McGraw Hill Publishing Company, Singapore.

Shotsberger, P., G., and Vetter, R. (2001), 'Teaching and Learning In The Wireless Classroom', ***IEEE Computer***, vol. 34, no 3, pp. 110-111.

Silberschatz, A., and Tuzhilin, A. (1995), 'On Objective Measures Of Interestingness In Knowledge Discovery', In: ***Proceedings Of KDD Annual Conference***, Montreal.

Silberschatz, A., and Tuzhilin, A. (1996), 'User Assisted Knowledge Discovery: How Much Should The User Be Involved', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec.

Silver, M. (1991), ***Systems That Support Decision Makers: Description And Analysis***, John Wiley and Sons, Inc., New York.

Singh, H. (1999), ***Interactive Data Warehousing***, Prentice-Hall, Inc., New Jersey.

Sizing Compaq Proliant Servers for Microsoft SQL Server 7.0 Data Marts. (1999), Compaq [online], URL: http://**www.compaq.com**

Sneath, P., H., A. (1957), 'The Application Of Computers To Taxonomy', ***J. Gen. Microbiology***, vol. 17, pp. 201-226.

Sneath, P., H., A. and Sokal, R. (1973), ***Numerical Taxonomy, Freeman***, San Francisco.

Sørensen, J., and Alnor, K. (1999), 'Creating Data Warehouse Using SQL Server', In: ***Proceedings Of The International Workshop On Design And Management Of Data Warehouses DMDW' 99***, Heidelberg, pp.10-1:10-9.

Spector, P. (1981), ***Research Design***, Sage Publications, London.

Sperley, E. (1999), ***The Enterprise Data Warehouse Planning, Building, And Implementation***, Prentice Hall PTR, Inc., New Jersey.

Sprague, H., and Carlson, E., D. (1982), ***Building Effective Decision Support Systems***, Prentice Hall, New Jersey.

Srivastava, J., and Chen, P. (1999), 'Warehouse Creation-A Potential Roadblock to Data Warehousing', ***IEEE Transactions On Knowledge And Data Engineering***, vol. 11, no 1, pp. 118-126.

Stair, R. (1992), ***Principles Of Information Systems- A Managerial Approach***, Boyd and Fraser Publishing company.

Staudt, M., Vaduva, A., and Vetterli, T. (2000), ***The Role of Metadata for Data Warehousing***, Swiss Federal Office of Professional Education and Technology.

Stevens, P. (1991), 'Decision Support Systems In Banking', In: ***Colloquium On Decision Support Systems***, The Institute of Electrical Engineers (IEE), London, pp. 3/1-3/4.

Stewart, B., Hetherington, G., and Smith, M. (1981), ***Survey Item Bank: Measures of Organisational Characteristics***, vol. 2, British Telecom, England.

Stowell, F., West, D., and Stansfield, M. (1997), 'Action Research As A Framework For IS Research', In Mingers, J., and Stowell, f., ***Information Systems: An Emerging Discipline?*** , McGraw-Hill, pp. 159-200.

Stroustrup, B. (1997), ***The C ++ Programming Language***, 3rd ed., Addison-Wesley, Massachusetts.

Swift, L. (1997), ***Mathematics and Statistics For Business, Management, and Finance***, Macmillan Press, London.

Taha, Y., Helal, A., and Ahmed, K. (1997), 'Data Warehousing: Usage, Architecture, And Research Issues', ***ISMM Microcomputer Application Journal***, vol. 16, no 2, pp. 70-80.

Teklitz, F., Krneta, P., and Puryear, R. (1999), SYBASE Business Intelligence on the Web with Windows NT, SYBASE BI [online], URL: http://**www.sybase.com/bi**

Teo, T., and King, W., R. (1996), 'Assessing The Impact Of Integrating Business Planning And IS Planning', ***Information & Management***, vol. 30, pp. 309-321.

The Egyptian Supreme Council Of Universities Statistics. (1999), The Egyptian Supreme Council of Universities, Cairo.

The IBM Business Intelligence Software Solution. (1999), IBM [online], URL: http://**www.dbaint.com**

The Ministry Of Higher Education Reports. (1980 to 1999), different volumes, *The Ministry Of Higher Education* Cairo.

Thearling, K. (1999), 'Increasing Customer Value By Integrating Data Mining And Campaign Management Software', ***Direct Marketing Magazine***, February.

Thierauf, R. (1988), ***User-Oriented Decision Support Systems: Accent On Problem***

Finding, Prentice-Hall, New Jersey.

Thomas, R. (2001), Overview of Canonical Variate Analysis (CVA), College Of Science, State University of Wayne [online], URL: **www.science.wayne.edu/~cwendorf/statman**

Travis, D. (1998), ***Thin Clients: Web-Based Client/Server Architecture And Applications***, McGraw-Hill.

Turban, E. (1993), ***Decision Support And Expert Systems, Management Support Systems***, Macmillan.

Turban, E., and Aronson, J. (1998), ***Decision Support Systems And Intelligent Systems***, 5th ed., Prentice Hall, New Jersey.

Urwiler, R., Ramarpu, N., Wilkes, R., and Frolick, M., N. (1995), 'Computer-Aided Software Engineering: The Determinants Of An Effective Implementation Strategy', ***Information & Management***, vol. 29, pp. 215-225.

Vandenbosch, B., and Huff, S. (1997), 'Searching Ad Scanning: How Executives Obtain Information From Executive Information Systems', ***MIS Quarterly***, vol. 21, no 1, pp. 81-107.

Volonino, L., and Watson, H. (1992), 'The Strategic Business Objectives Method For Guiding Executive Information Systems Development', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 145-159.

Waldroop, J., and Butler, T. (1996), 'The Executive As Coach', ***Harvard Business Review***, November-December, pp. 111-117.

Walker, M. (2001), Canonical Correlation Analysis, Unpublished Report, University of Colorado, Colorado [online], URL: **www.colorado.edu/epob/epob4640mwalker**

Watson, H. (1992), 'Avoiding Hidden EIS Pitfalls A Case Study: What You See Is Not Always What You Get', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 237-244.

Watson, H., and Frolick, M. (1992), 'Determining Information Requirements For An Executive Information System', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 161-176.

Watson, H., Rainer, R., and Koh, C. (1992), 'Executive Information Systems: A

Framework For Development And A Survey Of Current Practices', In Watson, H., Rainer, R., and Houdeshel, G. (Eds), ***Executive Information Systems: Emergence, Development, Impact***, John Wiley and Sons, Inc., New York, pp. 81-106.

Weyuker, E. (1986), 'Axiomatizing Software Test Data Adequacy', ***IEEE Transactions On Software Engineering***, vol. 12, no 12, pp. 1128-1138.

Whitten, J., Bentley, L., D., and Barlow, V., M. (1994), ***Systems Analysis And Design Methods***, 3rd ed., Irwin, Burr Ridge, Illinois.

Widom, J. (1995), 'Research Problems In Data Warehousing', In: ***Proceeding Of The 4th Int'l Conference On Information and Knowledge Management (CIKM) 95 ACM***, pp. 25-30.

Wiederhold, G. (1991), 'Views, Objects And Databases', ***Object-Oriented Database Systems***, Springer, pp. 29-43.

Wijsen, J. (2001), 'Trends in Databases: Reasoning and Mining', ***IEEE Transactions On Knowledge And Data Engineering***, vol. 13, no 3, pp. 426-438.

Willcocks, L., Feeny, D., and Islei, G. (1997), ***Managing IT as a Strategic Resource***, McGraw Hill, Berkshire.

Wilson, M. (1997), ***The Information Edge***, PITMAN Publishing, London.

Wishart, D. (1971), ***A General Approach To Cluster Analysis***, Part of Ph.D. Thesis, University Of St. Andrews.

Wishart, D. (1998), 'Efficient Hierarchical Cluster Analysis For Data Mining And Knowledge Discovery', ***Computing Science And Statistics***, vol. 30, pp. 257-263.

Wishart, D. (1999a), 'Clustan Graphics 3 Interactive Graphics For Cluster Analysis', In Gaul, W., and Locarek-Junge, H. (Eds), ***Classification In The Information Age***, Springer, pp. 268-275.

Wishart, D. (1999b), ***Clustan Graphics Primer: A Guide to Cluster Analysis***, Clustan Limited, Edinburgh.

Wishart, D. (2000), ***FocalPoint Clustering User Guide***, Clustan Limited, Edinburgh.

Wisniewski, E., Winston, H., Smith, R., and Kleyn, M. (1987), 'A Conceptual Clustering Program For Rule Generation', ***Int. J. Man-Machine Studies***, vol. 27, pp. 295-313.

Wixom, B. H., and Watson, H. J. (2001), 'An Empirical Investigation Of The Factors Affecting Data Warehousing Success', ***MIS Quarterly***, vol. 25, no 1, pp. 17-41.

Wonnacott, T., and Wonnacott, R., J. (1990), ***Introductory Statistics For Business and Economics***, 4th ed., John Wiley and Sons, Inc., Toronto.

World List of Universities and other Institutions of Higher Education. (2000), 22nd ed., Macmillan Reference Ltd., London.

Yossef, M. (1998), ***Web-Based Student Registration Systems***, unpublished MBA Thesis, The Advanced Management Institute (AMI), Alexandria.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996), 'BIRCH: An Efficient Data Clustering Method For Very Large Databases', In: ***Proceedings Of ACM SIGMOD '96 International Conference On Management Of Data***, June 4-6, Montreal, Quebec, pp. 103-114.

Zhang, X., and Rundensteiner, E., A. (1998), 'Data Warehouse Maintenance Under Concurrent Schema And Data Updates', ***Computer Science Technical Report Series***, August, pp. 1-29.

Zigurs, I., and Buckland, B. (1998), 'A Theory Of Task/Technology Fit And Group Support Systems Effectiveness', ***MIS Quarterly***, vol. 22, no 3, pp. 313-334.

Zikmund, W. (2000), ***Exploring Marketing Research***, 7th ed., The Dryden Press-Harcourt College Publishers, Orlando.

Zwass, V. (1998), ***Foundations Of Information Systems***, Irwin/McGraw-Hill, Boston, Massachusetts.

Glossary of terms used by the thesis

Activity Hierarchy Diagramming "AHD" (CASE tools' term)

The activity hierarchy diagram (AHD) identifies the lowest-level processes of interest to the business through decomposition. AHD shows levels of increasing detail for each function and process until activities decompose to the lowest level (elementary processes).

Ad hoc DSS

Deals with specific problems that are not recurring or frequent. For example the planning and budgeting decisions.

Additive

The most useful fact table should be additive. The reason for that is, every query runs against the fact table is expected to work on thousands or millions of records that require summarization and aggregations which are very hard to achieve if the fact table attributes are non-additive (i.e. non-additive attributes cannot be added at all).

Algorithm

An algorithm is a well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output.

Archival or historical data

When an organisation needs to store data about a specific topic for several years, it uses an archival or historical database. The archival database can contain either internal or external data sources or both.

Artificial Neural Networks (ANN)

An ANN are computer programs that implements complex pattern detection and machine learning algorithms to build predictive models from large database(s). In order for the ANN to detect patterns in the data sets, it should learn to detect these patterns and make predictions, in the same way a human does.

Association rules

Association rules are always defined on binary attributes. This sort of attributes makes it easy to describe records' profiles in any database.

Attribute/field/column

An attribute is a data item that identifies an entity/record in an entity type, e.g. name, address, prices, quantity...etc.

Business Area (CASE tools' term)

A business area names a collection of business functions and entity types to be analyzed

during a single analysis project.

Business Intelligence (BI)

BI is a popularized umbrella term coined in 1989 to describe a set of concepts and methods to improve business decision making by using DSS enhanced with any of these components; DW, OLAP, and data mining techniques.

Business System (CASE tools' term)

A business system is an integrated collection of procedures that support particular processes of a business. Business system components include the elementary processes specified in the Process Hierarchy Diagram, sequenced in the Process Dependency Diagram, and detailed in the action diagram.

Canonical Correlation

Canonical correlation is one of the multivariate analysis techniques that is used to study the interrelationships between two multiple variable sets; one set represents the independents (predictor variables) and the other set represents the dependents (criterion variables). This analysis can handle both variable types; the quantitative (metric) and the qualitative (nonmetric).

Centroid Cluster analysis

Groups are concentrated to lie in Euclidean space, and are then replaced by the co-ordinates of their centroids. The distance between groups is defined as the distance between their centroids.

Chi Square

Chi square test allows us to test for investigation of statistical significance in the analysis of a frequency distribution where H_o: $O_i = E_i$.

Clarity

Clear information means easy to understand

Classification

Classification is a learning function that classifies data records into one of several predefined classes.

Cleaning

Several methods are available to clean the data i.e. remove errors. Some of these methods can be executed in advance while others are only invoked after errors are detected at the coding or the discovery stage; duplication, domain inconsistency, and missing values.

Client/Server architecture

Client/Server architecture consists of a number of workstations (Clients), one or more higher configuration workstation(s) (Server(s)), and a local area network (LAN) connecting them all together. Client/server applications involve the dispersing of the software over several computers and creating a seamless environment for the end-users so that it appears as though they are working on just one system.

Clumping Clustering techniques

The previous groups of classification techniques end up with a number of disjoint clusters. However, in some other application areas the clusters need to be overlapping. Clumping techniques have been used in the area of language studies and disease diagnosis.

Cluster analysis

Clustering is basically classifying unclassified data. The data to be classified consists of a set of items (sometimes referred to as objects, fields, or records). Each item is described by a set of characteristics called variables (sometimes referred to as attributes). The target of clustering is to classify the items in the data set into a number of groups (sometimes referred to as classes, or clusters), such that objects within one group have similarities to one another.

Clustering

Clustering, Q-analysis, typology, grouping, clumping, classification, numerical taxonomy and unsupervised pattern recognition are used interchangeably to refer to the same thing. Clustering is the process of producing classifications from initially unclassified data.

Compound DSS

It is a hybrid system that includes two or more DSS types (e.g. text-oriented, rule-oriented, and solver-oriented). A compound DSS could be built by grouping a set of individual DSS each in one area of the decision situation.

Computer Aided Software Engineering (CASE) tools

CASE refers to the software used to assist in some or all phases of the SDLC. The basic feature of the CASE is the ability to generate an IS automatically based on the information stored about the business processes. CASE also offers a seamless transition from phase to phase with the same model throughout the SDLC.

Concurrent validity

To demonstrate concurrent validity, a test is correlated with another test of the same variable, both of which are administered concurrently.

Consolidation

Consolidation involves the aggregation of data, e.g. the total number of students at the university, total courses, and average GPA.

Construct

A construct is an image or idea specifically invented for a given research and/or theory-building purpose.

Construct validity

This attempts to identify the underlying construct(s) being measured and determine how well the test represents them, the extent a variable is abstract rather than concrete, we speak of it as being a construct.

Content validity

This measures the extent to which the instrument provides adequate coverage of the topic under study. Content validity requires a test to stand by itself as an adequate measure of what it supposed to measure.

Convergent validity

This attempts to demonstrate that each measure harmonizes with another measure. This could be done by using different measures to see how far there is convergence, or by using observations in addition to the questionnaire.

Correctness and completeness

Correctness means that information should be free from errors, whilst completeness refers to the degree to which information is free from omissions;

Critical Success Factors (CSF)

CSF are key areas of business activities, which if performed satisfactory will guarantee the success of the organisation.

Cronbach's Alpha Reliability

This calculates the average of all possible split-half reliability coefficients. Cronbach's Alpha is the most frequently used measure of internal consistency.

Data

Raw facts about things, events, activities, which are classified and recorded but not organized to convey a specific meaning

Data and Database heterogeneity

Data heterogeneity is the difference in how the data is defined in different database models. Database heterogeneity appears when the DW deals with different DBMS.

Data Mart

A smaller local data warehouse that is classified by subject is called a data mart.

Data miming

Searching for patterns in the data sets using analysis methods and models such as regression, clustering, SQL, visualization, decision trees and others.

Data Warehouse (DW)

A DW is a group of data extracted from different sources; internal, external, historical, and personal data archived in one or more data stores. The purpose of constructing a DW is to provide the DSS and the decision maker with the necessary data, which when transformed into information, will provide a better understanding of the business problem.

Database-Oriented DSS

In this type of DSS, the concentration is on the structure of the DSS itself. Normally database-oriented DSS are based on a relational structure, however, other database structure can be used such as object oriented or multi-media database structures.

Decision

A decision is a choice of one action among alternative course of actions.

Decision Support Systems (DSS)

A DSS is a computer-based information system that deals with semi-structured and unstructured problems facing managers at all management levels. The DSS goal is to enhance the decision quality and the manager effectiveness. To do so, the DSS integrates itself to the strategic data stores, which is the data warehouse (DW), and to the knowledge discovery in database (KDD) process that will find the deep knowledge and hidden patterns in the DW and present them to the DSS user.

Decision trees and rules

A decision tree is a predictive model that can be viewed as a tree. A decision tree is a means of visualizing complex decision problems where the questions can be posed in sequence.

Deep knowledge

Deep knowledge refers to the internal and causal structure of a system, and the interaction between its components relative to certain business application. The entire KDD process is able to handle this type of knowledge.

Dendrogram

The clustering process is done in two steps; *firstly* is to find out the similarity or distance between each pair of items, *secondly* based on the similarity/distance matrix a clustering

technique will be used to find out the number of groups and items within. Finally the groups found are depicted in a graph called *dendrogram.*

De-normalization

De-normalization helps to reduce the number of joins between the tables, thus making the query writing process easier, and also reduces the query execution time. Whereas the normalization process tries to split-up tables, the de-normalization process rolls-up all the data about the dimension in one table.

Density Search Clustering techniques

This group of clustering techniques assume that the items/records are distributed in two clustering areas one is a high-density area, and the other is of low density. Several techniques on this group are based on the NN and FN hierarchical techniques. This group of techniques has emerged to overcome the main problem of hierarchical techniques that is *chaining.*

Dependency modelling

It consists of finding a model that describes significant dependencies between variables. Dependency models fall into two categories structural and quantitative. The structural one determines which variables are dependent on each other, whilst the quantitative one shows the strength of the dependencies using a numerical scale.

Description

Description focuses on finding understandable patterns describing the data set. The relative importance of these goals varies from application to another, for some applications prediction is more important than description and for another applications description is more important.

Dimension table

The Star schema captures the measurements of importance to the business and the parameters by which the business measurements are broken down. It is a direct reflection on how business processes happen. The measurements are referred to as FACTS, whilst the parameters by which a measurement can be viewed are called DIMENSIONS.

Discriminant validity

This implies a low level of correspondence between a measure and other measures, which are supposed to represent other concepts. Discriminant validity is measured by Bivariate analysis.

Drill-down

This is the opposite of consolidation and involves more detailed inspection of the underlying data, e.g. the break down of the total number of students into different nationalities that belong to the different majors with different GPA.

DSS Generator

A computer software package that has the capabilities and facilities to enable the users to easily and rapidly build DSS, e.g. MS-Excel.

DW rollouts

Set of successive and phased tasks; each of which consists of a rollout number, expected users, requirements to be met, priority, and level of complexity.

Elementary Process (CASE tools' term)

An elementary process is any business activity that leaves the business in a consistent state, produces a result that is complete and meaningful to the user in itself, and occurs from beginning to end, at a single point in time in one place.

Elementary Process/Entity Type Matrix (CASE tools' term)

The elementary process/entity type matrix lets you record the expected effects of elementary processes on entity types. All activities have one of four effects on data: C = Create, R = Read, U = Update, D = Delete.

End-user computing

End-user computing, which is also known as end-user development, is the development of CBIS by people outside the formal information systems area e.g. managers, professional users (financial analyst, engineer, and lawyer). They build DSS to support their work and enhance their performance.

Enrichment

Enrichment is a method by which the analyst is able to have extra information about the data set(s) under study, for the purpose of enhancing the quality of the decision making process.

Enterprise DW

In some cases the DW can contain large number of fields and millions of data records about the entire organisation, and in such situations it is called an enterprise DW.

Entity type/table

An object of interest to the business, e.g. students, course, clients, products...etc.

Entity/record/row/occurrence

Set of related attributes.

Executive Information Systems (EIS)

EIS are highly interactive systems that provide managers and executives with flexible access to information for monitoring operating results and general business conditions, they are designed for the executives to use without any aid from intermediaries.

Executive Support Systems (ESS)

ESS are EIS with the capabilities of; supporting of electronic communications, providing data analysis capabilities, and facilitating organizing tools.

Exemplars

The exemplar is the most typical member of a cluster which has the minimum within-cluster-average based on a distance measure, or a maximum within-cluster-average based on a similarity measure.

Expert Systems (ES)

ES are computerized advisory systems that try to mimic the reasoning process and knowledge of experts in a specific domain.

External data

This is data that comes to the organisation from outside sources. There are many types e.g. government reports, federal publications, research institutions, commercial data banks, access to suppliers and customers' databases, and the Internet.

Face validity

This refers to the appearance of a test. Face validity concerns the acceptability of the test items; a test apparently reflects the contents of the concept(s) in its questions.

Fact table

The Star schema captures the measurements of importance to the business and the parameters by which the business measurements are broken down. It is a direct reflection on how business processes happen. The measurements are referred to as FACTS, whilst the parameters by which a measurement can be viewed are called DIMENSIONS.

Frequency

Information should be provided at an appropriate frequency to the decision maker

Front end of the DW

These are the tools used to analyze the data stored in the DW e.g. are general-purpose relational data access, data mining tools, DSS, EIS, and Web tools that perform search and query in the WWW environment.

Function (CASE tools' term)

A group of business activities that together completely support one aspect of furthering the mission of the enterprise. Each function describes something the business does, independent of the structure of the organization. In a function hierarchy (tree structure), the highest-level function is called the root function, and the lowest-level functions are called leaf functions.

Furthest neighbor (FN)

This technique is completely the opposite of the NN technique. Distance between groups is defined as the distance between the group and the farthest one. It ends with a diagram shows group fusions called *complete linkage dendrogram*. This technique falls in the same category of techniques as the NN (i.e. hierarchical techniques), uses the idea of group fusions, however, it does not always produce the same results that the NN produces for the same data sets.

Genetic Algorithms (GA)

Genetic algorithms loosely refer to these simulated evolutionary systems, but more precisely these are the algorithms that dictate how populations of organisms should be formed, evaluated and modified.

Global Executive Information Systems (GEIS)

GEIS are defined as being CBIS, provide access to internal and external data, used to support senior executives with analysis and decision making functions, and are only used by senior executives at headquarters in a global organisation.

Graphical User Interface (GUI)

GUI is a program interface, which relies on graphics capabilities to make the program easy to use. GUI uses a pointing device to select objects, icons, menus, graphics, and text boxes.

Group Decision Support Systems (GDSS)

When decisions are to be taken by a group, GDSS are used, GDSS allow a variety of specialists to be assembled whereby each of them is contributing to the solution using his expertise. GDSS could stimulate creative thinking and allow people from different departments to take the decision together.

Group support

The support is given to a group of people who are engaged in separate decisions but the decisions are correlated.

Hidden knowledge

This knowledge can be found by using groups of the data mining techniques, such as ANN

& OLAP or GA & decision trees. Combining more than one data mining technique for the same data set should take into consideration the match between the type of this data and the characteristics of these techniques.

Hierarchical

A hierarchy (or tree) is a network in which nodes are connected by links such that all links point in the direction from child to parent. Each node has one parent and there is always one path between any two nodes. The hierarchical model stores the data fields in a top-down order. The database looks like a tree, and there are links between related fields. The basic operation in a hierarchy is the tree search, when a query is processed the nodes that meet the conditions of the query will be returned.

Hybrid Support Systems

All computer-based information systems (CBIS) share the same objective that is to assist managers in their decision making, or in other words to transfer the managers from an uncertain situation into a situation of certainty or risk (any point between certainty and uncertainty) by providing complete or partially complete information. To complete this process one or more information systems might be used.

Indexing

In order to reduce the processing time the DW designer makes extensive use of indexing. An index functions like a smaller table in ordered sequence and provides direct access to the rows of interest to a user without having to scan all the rows of the entire table. To ensure an optimal indexing methodology, multiple indexes are created on most of the dimension tables.

Information

Information is data that has been processed and has specific meaning to someone

Information Systems (IS)

IS can be defined as a set of interrelated components working together to collect, retrieve, process, store, and disseminate information for the purpose of facilitating planning, control, coordination, and decision making in business and other organisations. These components are people, organisations, and technology.

Institutional DSS

This type of DSS deals with the routine and frequent problems in organisations. Examples include evaluating investment opportunities, which may be built incrementally across years.

Integrated DW

In many organisations the same piece of data may exist on several databases, to overcome the data redundancy problem there has to be an integration of data sources to avoid duplication.

Internal data

These are the data sources of an organisation that cover the whole business, e.g. data about employees, daily transactions, products, stock levels, customers etc. Usually internal data is stored in one or more databases; they also created by organisations when using TPS.

Knowledge

A combination of experience, accumulated learning, and information that have been organized and analyzed to be understandable and applicable to a specific decision situation

Knowledge discovery in database (KDD)

KDD is defined as the process of finding patterns of hidden information or unknown facts in the database. Traditionally the notion of finding useful unknown patterns and hidden information in raw data has been given many titles including knowledge discovery in database, data mining, data archaeology, information discovery, knowledge discovery or extraction, and information harvesting.

Knowledge-based decision support systems (KBDSS)

KBDSS have been developed, however few systems address the use of knowledge in the decision problem. KBDSS are DSS enhanced with a knowledge component.

Management Information Systems (MIS)

MIS have emerged in response to the shortcomings of TPS in order to provide the information required for managing the organisation.

Model base

The model base provides the analysis capability of the DSS. It is important to realize that the DSS can contain one model in some situations and up to several hundreds in others.

Model Base Management Systems (MBMS)

The functions of MBMS are model creation, updating and model data manipulation.

Model directory

The function of the model directory is similar to that of the data directory. The model directory contains the model description, its main functions, and its capabilities.

Model execution, integration, and command processor

Model execution is used to control the model whilst it is in use. Model integration is the

process of integrating one or more models in one problem. And a model command processor is utilized to accept and interpret instructions and send them to the MBMS.

Modelling language

Modelling languages are utilized to write the DSS models using general-purpose languages like COBOL, BASIC, or special modelling languages like IFPS (interactive financial planning system).

Multi-dimensional knowledge

This type of knowledge can only be obtained using some specific techniques such as OLAP and visualization. For instance, the OLAP technique is used to make the output multi-dimensional.

Multi-media

The multi-media model manages data in many formats; text, numeric, images, bit-maps, pictures, hypertext, video clips, sounds and multi-dimensional images (virtual reality).

Multi-media database management system (MMDBMS)

The software that utilizes the multi-media model in creating and maintaining databases is called a MMDBMS.

Nearest neighbor (NN)

After formulating the proximity matrix the smallest item is found and the two items are then fused together to formulate a new item, each fusion decreases the number of items by one. The distance between groups is defined as the distance between their closest members. This NN technique ends with a diagram called *single linkage dendrogram* showing group fusions.

Network

The network model sometimes called the CODASYL model, uses additional pointers to give the hierarchical model more flexibility. The network model allows more complex links between nodes. Thus the hierarchical model may be viewed as a special case of the network model where each node is linked to a parent node only. The network model saves storage space through the sharing of data items.

Non-volatile DW

The objective of using the DW is to respond to management requests for information. This data is extracted from the operational database and then loaded into the DW database. This means that a data warehouse will always be filled with historical data and should be updated regularly from the operational database.

Normalization

It is the process eliminating redundancy from entity types. It includes the removal of redundant: attributes, keys, and relationships from the conceptual data model.

Object-oriented

This model is used with complex applications, which require accessibility to data that have complex and inter-related relationships, e.g. computer aided design and manufacturing (CAD/CAM), computer integrated manufacturing (CIM), and geographic information systems (GIS).

Object-oriented database management systems (OODBMS)

The software package that utilizes the object oriented model in creating and maintaining databases is called OODBMS. It allows the analysis of the DB in terms of objects. Abstraction is used to develop the inheritance between object levels, encapsulation allows the DB designer to store conventional and procedural code within the same objects.

Office Automation Systems (OAS)

OAS refer to any IS associated with general office work or office activities. The activities performed by office staff in an organisation include; managing documents, scheduling individuals and groups, managing data, and managing projects.

On Line Analytical Processing (OLAP)

The key driver for the development of OLAP is to enable the multi-dimensional access to these analysis tools; consolidation, drill-down, and slicing and dicing.

Operational Data Stores (ODS)

Organisations use TPS to store their business transactions. These TPS have a certain database design requirements and for this reason these types of databases are known as ODS, or on-line transaction processing applications (OLTP). They are optimized to handle large numbers of concurrent users either storing/editing transactions or processing queries.

Operational models

Operational models are utilized to support the analysis of routine and frequent problems. These models normally use internal data sources. The first-line managers are the fundamental users of these models.

Optimization clustering techniques

Optimization techniques apply a relocation of the items/records which allows that a wrong-initial partition to be corrected at a later stage. Some of the optimization techniques allow the number of groups to be changed later in the analysis. Problems associated with these

techniques emerge when deciding on the number of clusters to be retained.

Organisational support

This DSS type focuses on the organisational tasks in a sequence of operations or different location and resources.

Parallel/Alternative-forms Reliability

This involves administering the same test in an alternative or parallel form to the same person simultaneously or with a small delay. Here we have two versions of the same test linked in a systematic way. Each person is given the two tests to complete and the reliability is obtained by calculating the Pearson Correlation coefficient.

Personal data

This data source includes the manager's own experience and opinions and/or estimates about market share, additional customer data or other policies. These personal data sources are used in EIS and DSS.

Personal support

The support here is given to an individual taking a decision.

Prediction

Prediction is the use of some variables or data fields in the database to predict the unknown future values of the other variables or data fields of interest.

Predictive validity/Criterion-related validity

A test is said to possess predictive validity if it can predict some relevant outcome. Predictive validity is concerned when the purpose of an instrument is to estimate some variable that is external to it, which is referred to as the criterion.

Primary Key (PK)

A PK is unique identifier; no entity has more than one PK value, and one PK value is assigned to one entity.

Prototyping

Prototyping builds the DSS through a set of steps with spontaneous feedback from the user manager to ensure that the development process is running on the proper track. The prototyping is sometimes called the evolutionary approach or the iterative process of just prototyping.

Proximity measures

There are many methods to find the similarity or distance between items; these methods are called *proximity measures*. All proximity measures end up with either a *similarity*

matrix (if it is a similarity measure) or a *distance matrix* (if it is a distance measure).

Regression

A regression model is a mathematical equation that provides predictions of the values of one variable (dependent) based on the known values of one or more other variables (independent or predictors).

Relationship

An association between entity types (i.e. tables), e.g. students and courses, accounts and clients...etc.

Relevance

Information should be provided to the relevant person

Reliability

The reliability of a measure refers to its consistency. A measurement is reliable to the degree that it supplies consistent results.

Rule

A rule is a method of defining DSS directions/instructions, expressed in the format of IF ... THEN... statements.

Rule-oriented DSS

The knowledge component of the DSS includes rules; these rules are either qualitative or quantitative. Rules are the principal components of the knowledge base DSS; it extends the capabilities of the computer far beyond the data or model-base.

Scalability

Scalability is the ability to scale the hardware and/or software to support larger or smaller volumes of data and more or less users.

Semi-structured decisions

That is, some decision procedures can be pre-specified, but not enough to lead to a definite recommended decision. For example, decisions involved in starting a new line of products or making a major change to employee benefits would probably range from unstructured to semi-structured.

Shallow knowledge

Shallow knowledge refers to the representation of surface level knowledge, using the input/output relationship of the system, and hence can be represented in the format of IF-THEN rules.

Slicing and dicing

Slicing and dicing refers to the ability to look at the database from different viewpoints.

Solver-oriented DSS

A solver is a technique or computer program written to resolve a certain computation or a particular problem. Examples are the reorder level in a stock control system, optimum process settings, etc.

Split-half Reliability

This establishes the degree to which instrument items are homogenous and reflect the same underlying construct(s). The items in a scale are divided into two groups either randomly or on an odd-even basis, and the relationship between respondents' scores for the two halves is computed.

Spreadsheet-oriented DSS

A spreadsheet is a modelling language that allows the user to write models to execute the DSS analysis. Spreadsheets are widely used by end-users; the most common examples are Microsoft Excel, Lotus 123, and Q-Pro.

Star Schema Structure

The best way to build the DW database is by using the star schema structure (sometimes referred to as multi-dimensional data modelling-MDDM). A simple star consists of group of tables that describe the dimensions of the business arranged around a central table that contains the business facts. The smaller outer tables are the points of the star, the larger table in the center is the star from which the points radiate. The star schema relies on two major components the facts and the dimensions.

Strategic alliance

A strategic alliance is a mutually dependent relationship where the success or failure of one party affects the other. By which businesses are able to share and exchange data.

Strategic Information Systems (SIS)

Any information system that performs a strategic role, which involves the development of products and services and capabilities that give the organisation strategic advantages over the strategic forces in the market is called a strategic information system, or a strategic management information systems (SMIS). SIS can be any kind of IS, it might be TPS, MIS, EIS, or DSS.

Strategic models

Strategic models tend to be broader than the other models in spectrum and embody many variables and often use external as well as internal data.

Structured decisions

Involve situations where the procedures to follow when a decision is needed can be specified in advance. The inventory reorder decisions faced by most businesses are a typical example.

Structured Query Language (SQL)

SQL tools are used to extract data that matches search criteria or represent this data in a way the user finds it easier to handle or interpret.

Subject Area (CASE tools' term)

A subject area is a cohesive group of entity types (with their relationships, attributes and operations).

Subject-oriented DW

Data are organized according to subject instead of application. Examples of subjects are marketing, production, personnel, sales etc.

Summarization

It is the process of finding a compact description for a subset of data. Summarization is often applied to interactive exploratory data analysis and report generation.

Summary tables

A summary table is a DW table that includes data frequently retrieved by users. Instead of searching in the entire fact table a snapshot is taken and stored in a summary table, when the user invokes the relevant query the result comes from the summary table.

Tactical models

Tactical models are used by middle level managers to assist them in the resource allocation and control related problems. Some external data may be required but the main data source is internal.

Test-retest Reliability

This is used to measure the external reliability by administering the same test twice to the same subjects over an interval of less than six months to ensure the stability of results.

Text-oriented DSS

Decision makers should be able to access the corporate stored databases that are always in a textual format. A text-oriented DSS helps the decision maker by allowing the document to be electronically created, revised, indexed, and processed as needed.

The Canonical Correlation analysis

Canonical correlation is one of the multivariate analysis techniques. It is used to study the

interrelationships between two multiple variable sets; one set represents the independents (predictor variables) and the other set represents the dependents (criterion variables). This analysis can handle both variable types; the quantitative (metric) and the qualitative (nonmetric).

The Chi Square Test

Chi Square is a non-parametric distribution-free test that is used with nominal and ordinal data. Distribution free means that these tests are free of assumptions regarding the data distribution of the parent population.

The data directory

The data directory or as sometimes called data dictionary or catalog which includes data about data; it includes the data definitions and its functions. The directory supports the data maintenance function. It is a central location of metadata.

The Database (DB)

A DB is a collection of related data that have a common purpose.

The DBMS

The DBMS is a software package that is used to create and maintain a database

The decision maker

The decision maker is the one who uses the DSS as a tool to enhance his information about the situation(s).

The Granularity of the Fact table

The term granularity (sometimes referred to as grain of the fact table) describes the level of detail stored in the fact table and follows the level of detail of its related dimensions.

The information engineering (IE) approach

The basic idea of the IE approach is to bring the organisation's plans into the process of IS development. That is each IS development is derived from a certain business requirements; the requirements are based on the organisation goals and objectives. These goals and objectives and the business requirements drive the information systems development plans.

The Information Need/Objective Matrix (CASE tools' term)

The information need/objective matrix maps long-term objectives to the information needs required to meet the objectives. This matrix highlights the information that is required to accomplish long-term objectives. You can use the information to assign priorities to the development of information systems.

The multi-tier DW

The multi-tier (3-tier) DW architecture or as sometimes called the thin client model, handles the scalability and flexibility problems through the application servers. Application servers perform data filtering, summarization, aggregation, support meta data, data access, and provides multi-dimensional views.

The query facility

The query facility using its query language accepts requests from the other DSS components and returns the required results.

The relational model

This is the most frequently used database model within the DBMS context. It is also the dominant database model in DSS applications, and frequently used in the development of a DW. It allows the user to think of the DB model in terms of two-dimensional tables. A table consists of rows and columns, rows are the data records and columns are the individual fields. Data tables are joined to each other's be creating relationships.

The Systems Development Life cycle (SDLC)

The SDLC is a process by which systems analysts, software engineers, programmers, and end-users build information systems. SDLC may be considered to contain eight steps; planning, research, system analysis and conceptual design, design, construction, implementation and user training, maintenance, and adaptation.

The Top-down DW development approach

In the top down approach, an organisation need to develop an enterprise data model, collect enterprise-wide business requirements and then builds an enterprise data warehouse with subset data marts.

The Top-down DW development approach

In the bottom up approach, an organisation needs to prioritize the development of individual data marts, which are then integrated into the enterprise data warehouse.

The two-tier DW

The two-tier (2-tier) DW architecture or as sometimes called the fat client model, in which clients' functions include GUI presentation logic, query definition, data analysis, report formatting, summarization, and data access, whilst the DW server performs data logic, data services, metadata maintenance and the file services.

The user interface

The user interface is the medium between the user of the DSS and the DSS itself, it encompasses the manager's preferences.

Timeliness

It means that the information is being provided to the right person at the right time.

Time-variant DW

The DW contains data gathered from different periods. The DW contains a place for storing historical data that can be used for comparisons, trends, or forecasting. Historical data can be over twenty years old.

Transaction

A transaction is a business event that generates or modifies data. Each single transaction represents a single business event.

Transaction Processing Systems (TPS)

TPS record and collect data about the daily transactions taking place in any organisation.

Unstructured decisions

Involve decision situations where it is not possible to specify in advance most of the decision procedures to follow.

Validity

Validity refers to the extent to which a test measures the concept(s) that it intends or claims to measure.

Value chain

Managers inside the organisation try to use every weapon they have to increase the value of their products/services, and this what is called the value chain

Visualization

Visualization refers to presenting data and summary information in graphics; it depends strongly on the human side of the analysis. It is an emerging technology that allows organizations to process information and present it in a usable format.

Ward's method

This technique assumes that group fusions results in a loss of information at each stage of the analysis. This information loss can be measured by the total sum of squared deviations for every point from its mean of the cluster it belongs to. Each stage fusions are done for those who have the minimum increase in error sum of squares.

Appendices

Appendix (A)

The ARDSSQ

Research Questionnaire

University of Plymouth
Business School

Arab Academy for Science and Technology (AAST)
College of Management & Technology

Admission & Registration Decision Support System Questionnaire

Dear respondent

This questionnaire is about the admission and registration functions in universities. One objective of the questionnaire is to define the current admission and registration information systems and to describe their features. Another objective is to explore the requirements that are not satisfied by the current systems.

To meet these objectives, the questionnaire is divided into two parts. The first part requests information about your current admission and registration information system. The second part of the questionnaire asks you to identify your information needs.

For the sole purpose of research, you are kindly requested to accurately and objectively answer all of the questions. Please be advised that all answers will be kept confidential.

When you have completed the questionnaire please return it to the researcher in the enclosed envelope.

Thank you in advance for your cooperation,
The researcher,

Ahmed Abdel Hamid El-Ragal
Management Information Systems Department, AAST
P.O. Box 1029, Miami, Alexandria, Egypt

-Note:
Please if you have any inquiries or would like further information about the questionnaire please call one of these telephones: 010-5111600, 03-5503220, or you can send e-mail to: a.elragal@computer.org

General Information

- **Please kindly complete the following information:**

-University data

1. Name of the University: .....
2. Type of University : ✓

Government	
Private	

3. College/Faculty/Higher Institute: .....
4. Address(Optional) : .....

.....

.....

-Respondent data

5. Name (Optional) : .....
6. Position : ✓

Dean	
Deputy Dean	
Registrar	
Admission Officer	
Other	

7. Tel (Optional) : Office ..... Mobile .....
8. E-mail (Optional) : .....

Part I- Current Admission and Registration Information System

- **Please answer the following Yes/No questions by selecting only one choice using the sign (✓). Remember that there are no right or wrong answers:**

1. Do you believe that the manager who is responsible for admission and registration related decisions in universities should have a computer on his/her desk?

Yes	
No	

2. Do you have a computer on your desk?

Yes	
No	

3. Do you currently use a computerized admission and registration information system?

Yes	
No	

*If your answer is **No**, please go to part II Information needs, if **Yes** proceed to question no.4*

4. Do you use the admission and registration information system to perform all of your admission and registration functions?

Yes	
No	

5. Is your information system linked to an archival or historical students' database?

Yes	
No	

6. Do you depend on the current information system to take decisions?

Yes	
No	

*If your answer is **No**, please go to question no.8, if **Yes** proceed to question no.7*

7. Do you encounter situations where your decision will be enhanced if you search in the students' history before making the decision?

Yes	
No	

8. If the output you received from a query made on the admission and registration information system was unexpected and/or surprising. Will you use the system result and update your experience?

Yes	
No	

*If your answer is **Yes**, please go to question no.11, if No proceed to question no.9*

9. If the output you received from a query made on the admission and registration information system was unexpected and/or surprising. Will you use your own experience and discard the system result?

Yes	
No	

*If your answer is **Yes**, please go to question no.11, if No proceed to question no.10*

10. If the output you received from a query made on the admission and registration information system was unexpected and/or surprising. Will you use a combination of the system result and own experience depending on the situation?

Yes	
No	

- **Please answer the following questions by choosing all that apply using the sign (✓). Remember that there are no right or wrong answers:**

11. Among the following admission and registration information system features, choose **Yes** if your system includes the feature or **No** it the feature is not available.

	Yes	**No**
Printing reports that describe students' records		
Predicting the new applicants' performance		
Predicting the current-students' performance		
Both description and prediction functions are available		
It is an electronic store of students' data		
Other		

12. Among the following admission and registration information system functions, choose **Yes** if the function is part of your system or **No** it is not part of it.

	Yes	No
Student description reports e.g. *Transcripts*		
Student performance prediction		
General statistics		
Classifying students into similar groups		
Finding relationships between a student's data fields e.g. *The relationship between age and Grade Point Average (GPA)*		
Using the historical data to describe the Students' history		
Using external data to enhance the quality of decisions		
Other		

Part II- Information needs

- **Please answer the following (Yes/No) questions, using the sign (✓). Remember that there are no right or wrong answers:**

13. I believe that the main role of computer is electronic data storage.

Yes	
No	

14. I believe that one of the computer roles is to be a decision maker.

Yes	
No	

15. The higher the rank of the decision maker is in the chain of command, the reports produced by the system are required to contain more detail.

Yes	
No	

16. The fact that my competitors-outside the organization- may have access to the same information makes it less useful.

Yes	
No	

17. Your ideal admission and registration information system would provide information from internal data sources.

Yes	
No	

18. Your ideal admission and registration information system would provide information from external data sources.

Yes	
No	

19. The admission and registration information system should be able to forecast in addition to providing summary statistics.

Yes	
No	

20. I believe that admission and registration information systems would be able to improve the quality of the managers' decisions.

Yes	
No	

21. The admission and registration information system should be able to help managers take decisions.

Yes	
No	

22. The more actionable the result gained from the admission and registration information system, the more accepted is this result.

Yes	
No	

---

23. The following is a list of decisions related to the admission and registration functions in universities. Assuming that you have a computerized admission and registration information system that is capable of taking decisions, please choose **Yes** for those decisions that the system should take, or **No** for those that it should not. Use the sign (✓) to make your choices. You can use the space provided at the end of this table to add any other decision(s). Remember there are no right or wrong answers.

No.	Decision	Yes	No
• **Admission-related decisions:**			
A.	Accept or reject a new applicant		
B.	Provide unconditional offer for new applicant		
C.	Predict the new applicants that will join the faculty/college/institute this term/year based on our archival records		
D.	Predict the new applicants that will join the college this term/year based on government statistics on secondary school students		
E.	Predict the new applicants that will join the college this term/year based on our archival records besides other records like the government statistics		
F.	Based on our archival records we can make an applicant-major match and provide this to the new applicant to help him/her chooses a suitable major		
G.	Hold the applicant until the following term/year		
H.	Accept or reject the applicant who is transferred from another educational institution		
I.	Accept or reject the applicant who is transferred from another educational institution based on our transfer history records		
• **Registration-related decisions:**			
J.	Predict a student's performance based on the students' history we keep		
K.	Predict a course's results based on the courses' history we keep		
L.	Classifying students into similar groups		
M.	Predict a student's performance based on the group that he/she belongs to		
N.	Set the student status to "On probation"		
O.	Predict the "On probation" students based on the students' history we keep		
P.	Make relationships between students' performance and academic departments		
Q.	Forecast course booking		
R.	Decide on Student abandonment		
• **Other decisions you think should be taken by the system:**			
S.	....		

T.	.....
U.	.....
V.	.....
W.	.....

24. Given the following list of admission and registration functions, please choose **Yes** for those functions that should be part of the ideal admission and registration information system, or **No** for those that should not be part of it. Use the sign (✓) to make your choice. Remember that there are no right or wrong answers:

No.	Function name	Yes	No
A.	New applicants' performance prediction		
B.	Current-students' performance prediction		
C.	Student description reports e.g. *transcripts*		
D.	General statistics		
E.	Classifying students into groups		
F.	Using historical data		
G.	Using external data e.g. *government regulations*		
H.	Finding relationships between students' data fields		
I.	Creating ad hoc reports that are not structured on the admission and registration information system		
J.	Others		
K.	Others		

25. Given the following characteristics of information systems, please choose **Yes** for those characteristics that should be part of the ideal admission and registration information system, or **No** for those that should not be part of it. Use the sign (✓) to make your choice. Remember that there are no right or wrong answers:

No.	Characteristic name	Yes	No
A.	Ease of use		
B.	Requires minimum training time to be learned		
C.	User-involvement during the system development process		
D.	Design the system to be able to grow in future		
E.	Flexible system		
F.	Can be integrated with other systems		
G.	E-mail facility		
H.	Accessible through the Web		
I.	Cost-effective		
J.	Others		
K.	Others		

Thank you for your time and cooperation

Please return your completed questionnaire in the pre-paid envelope supplied

الأكاديمية العربية للعلوم و التكنولوجيا– مصر
كلية الإدارة والتكنولوجيا

جامعة بليموث – إنجلترا
كلية إدارة الأعمال

إستقصاء عن نظم دعم قرارات القبول و التسجيل
Admission & Registration Decision Support System Questionnaire

عزيزي المستجيب

يختص هذا الإستقصاء بوظيفة القبول و التسجيل في الجامعات. الهدف الأول من هذا الإستقصاء هو وصف أنظمة معلومات القبول و التسجيل الحالية و التعرف علي خصائصها. أما الهدف الثاني فهو التعرف علي الإحتياجات من المعلومات و التي لا تستطيع الأنظمة الحالية الوفاء بها.

لتحقيق هذان الهدفان تم تقسيم الإستقصاء إلي جزئين. الجزء الأول يختص بجمع معلومات عن نظام القبول و التسجيل الذي تستخدمه حاليا. أما الجزء الثاني يختص بجمع معلومات عن احتياجاتك من المعلومات.

لأغراض البحث العلمي فقط، من فضلك أجب بدقة و موضوعية عن جميع الأسئلة. نرجو العلم بأن جميع الإجابات ستكون سرّية.

بعد أن تقوم بملأ هذا الإستقصاء من فضلك أرسله للباحث مستخدما المظروف المرسل والملصق عليه طابع بريد.

شكرا جزيلا علي تعاونكم،

الباحث

أحمد عبد الحميد الرّحال
كلية الإدارة و التكنولوجيا- الأكاديمية العربية للعلوم و التكنولوجيا
ص.ب. ١٠٢٩ ميامي - الإسكندرية - مصر

ملاحظة:

في حالة وجود أي استفسار بخصوص هذا الإستقصاء من فضلك أتصل بأحد الأرقام التالية: ٠١٠٥١١١٦٠٠ أو ٠٣٥٥٠٣٢٢٠ ، أو أرسل بريد إلكتروني إلي: aelragal@mis.aast.edu

معلومات عامة

- **من فضلك أجب علي الأسئلة التالية:**

- بيانات عن الجامعة

١. إسم الجامعة .....

٢. نوع الجامعة

	✓
حكومية	
خاصة	

٣. إسم الكلية \ المعهد .....

٤. العنوان (اختياري) .....

.....

.....

- بيانات عن المستجيب

٥. الإسم (اختياري) .....

٦. الوظيفة

	✓
العميد	
وكيل الكلية لشئون الطلاب	
المسجل	
مسئول القبول	
أخري	

٧. التليفون (اختياري) عمل ..... محمول .....

٨. البريد الإلكتروني (اختياري) .....

الجزء الأول: نظام معلومات القبول و التسجيل الحالي

- **من فضلك أجب علي أسئلة (نعم \ لا) التالية مستخدما العلامة (✓). تذكر أنه لا توجد إجابات صحيحة و إجابات خاطئة:**

١. هل تعتقد بأن المدير المسئول عن إتخاذ القرارات المرتبطة بعملية القبول و التسجيل في الجامعات يجب أن يكون لديه جهاز حاسب آلي؟

نعم	
لا	

٢. هل لديك جهاز حاسب آلي في مكتبك؟

نعم	
لا	

٣. هل تستخدم حاليا نظام معلومات آلي للقبول و التسجيل؟

نعم	
لا	

إذا كانت إجابتك (**لا**) من فضلك إنتقل للجزء الثاني: **الإحتياجات من المعلومات**، أما لو كانت الإجابة (**نعم**) فإنتقل إلي السؤال التالي (٤)

٤. هل تستخدم نظام معلومات القبول و التسجيل لأداء جميع مهامك الوظيفية من خلاله؟

نعم	
لا	

٥. هل نظام المعلومات الذي تستخدمه مرتبط بالسجلات التاريخية التي تحتفظون بها؟

نعم	
لا	

٦. هل تعتمد علي نظام المعلومات الحالي في إتخاذ القرارات؟

نعم	
لا	

إذا كانت إجابتك (**لا**) من فضلك إنتقل للسؤال رقم (٨)، أما لو كانت الإجابة (**نعم**) فإنتقل إلي السؤال التالي (٧)

٧. هل تواجه بمواقف يكون القرار المزمع اتخاذه فيها يمكن تحسينه من خلال البحث في السجلات التاريخية قبل إتخاذ هذا القرار؟

نعم	
لا	

٨. إذا كانت النتيجة التي حصلت عليها من أحد الإستفسارات query التي أجريتها علي نظام معلومات القبول و التسجيل غير متوقعة أو مفاجئة لك. هل تتبع النتائج المستخرجة من النظام و تقوم بتحديث خبراتك و معلوماتك؟

نعم	
لا	

إذا كانت إجابتك (**نعم**) من فضلك إنتقل للسؤال رقم (١١)، أما لو كانت الإجابة (**لا**) فإنتقل إلي السؤال التالي (٩)

٩. إذا كانت النتيجة التي حصلت عليها من أحد الإستفسارات query التي أجريتها علي نظام معلومات القبول و التسجيل غير متوقعة أو مفاجئة لك. هل تتبع خبراتك و معلوماتك و تتجاهل النتيجة المستخرجة من النظام؟

نعم	
لا	

إذا كانت إجابتك (**نعم**) من فضلك إنتقل للسؤال رقم (١١)، أما لو كانت الإجابة (**لا**) فإنتقل إلي السؤال التالي (١٠)

١٠. إذا كانت النتيجة التي حصلت عليها من أحد الإستفسارات query التي أجريتها علي نظام معلومات القبول و التسجيل غير متوقعة أو مفاجئة لك. هل تستخدم مزيج من خبراتك و نتائج النظام في نسب معينة تعتمد علي الموقف أو القرار المزمع اتخاذه؟

نعم	
لا	

- **من فضلك أجب علي الأسئلة التالية مستخدما العلامة (✓) . يمكنك إختيار أكثر من عنصر واحد في نفس السؤال. تذكر أنه لا توجد إجابات صحيحة و إجابات خاطئة:**

١١. الجدول التالي يحتوي علي مجموعة من خصائص نظم معلومات القبول و التسجيل. إختار (**نعم**) للخصائص الموجودة في نظام معلومات القبول و التسجيل الذي تستخدمه حاليا، أو (**لا**) للخصائص الغير متاحة.

	نعم	**لا**
طباعة تقارير تصف سجلات الطلاب		
التنبؤ بالأداء العلمي للطلبة الجدد		
التنبؤ بالأداء العلمي للطلبة القدامى		
كل من خاصيتي الوصف و التنبؤ متاحة		
هو عبارة عن مخزن إلكتروني لبيانات الطلاب		
أخرى		

١٢. الجدول التالي يحتوي علي مجموعة من وظائف نظم معلومات القبول و التسجيل. إختار (**نعم**) للوظائف الموجودة في نظام معلومات القبول و التسجيل الذي تستخدمه حاليا، أو (**لا**) للوظائف الغير موجودة.

	نعم	لا
تقارير تصف حالة الطلاب العلمية مثال: شهادة بالإنجاز transcript		
التنبؤ بالأداء العلمي للطلاب		
إحصائيات عامة		
تقسيم الطلاب إلي مجموعات متشابهة		
إيجاد علاقات بين بيانات الطلاب المختلفة مثال: العلاقة بين العمر و التقدير أو المعدل		
إستخدام البيانات التاريخية لتصنيف الطلاب		
إستخدام البيانات الخارجية لتحسين جودة القرارات		
أخري		

الجزء الثاني: الإحتياجات من المعلومات

- **من فضلك أجب علي أسئلة (نعم \ لا) التالية مستخدما العلامة (✓). تذكر أنه لا توجد إجابات صحيحة و إجابات خاطئة:**

١٣. أعتقد أن الدور الرئيسي للحاسب الآلي هو أنه وعاء إلكتروني لتخزين البيانات.

نعم	
لا	

١٤. أعتقد أن أحد أدوار الحاسب الآلي هو إتخاذ القرارات.

نعم	
لا	

١٥. كلما ارتفعت مكانة متخذ القرار في الهيكل التنظيمي كلما كانت التقارير التي تصل إليه من النظام تحتوي علي كم أكبر من التفاصيل.

نعم	
لا	

١٦. إن توافر نفس المعلومات للمنافسين (من خارج المنظمة التي تعمل بها) قد يجعل هذه المعلومات أقل فائدة.

نعم	
لا	

١٧. إن نظام معلومات القبول و التسجيل المثالي يجب أن يقدم معلومات معتمدة علي مصادر البيانات الداخلية.

نعم	
لا	

١٨. إن نظام معلومات القبول و التسجيل المثالي يجب أن يقدم معلومات معتمدة علي مصادر البيانات الخارجية.

نعم	
لا	

١٩. إن نظام معلومات القبول و التسجيل يجب أن تكون لديه القدرة علي التنبؤ بالإضافة إلي تقديم الإحصائيات.

نعم	
لا	

٢٠. أعتقد بأن نظام معلومات القبول و التسجيل لديه القدرة علي تحسين جودة القرارات التي يتخذها المدير.

نعم	
لا	

٢١. إن نظام معلومات القبول و التسجيل يجب أن يساعد المدير في عملية اتخاذ القرارات.

نعم	
لا	

٢٢. كلما كانت النتائج المستخلصة من نظام معلومات القبول و التسجيل يمكن تطبيقها كلما إرتفعت نسبة قبول هذه النتائج.

نعم	
لا	

---

٢٣. الجدول التالي يحتوي علي بعض القرارات المرتبطة بعملية القبول و التسجيل في الجامعات. علي فرض أن لديك نظام معلومات للقبول و التسجيل لديه القدرة علي إتخاذ القرارات. من فضلك إختار (نعم) للقرارات التي يجب أن يتخذها النظام، أو (لا) للقرارات التي لا يجب أن يقوم النظام باتخاذها. أستخدم العلامة (✓) لعمل اختياراتك. من الممكن أن تستخدم المساحة الخالية أسفل الجدول لإضافة قرارات جديدة. تذكر أنه لا توجد إجابات صحيحة و إجابات خاطئة:

رقم	القرار	نعم	لا
•	**قرارات مرتبطة بعملية القبول:**		
أ.	قبول أو رفض الطالب المستجد		
ب.	قبول غير مشروط للطالب المستجد		
ت.	التنبؤ بأعداد الطلاب المستجدين المتوقع التحاقهم بكل كلية\معهد للفصل\العام القادم بناء علي السجلات التاريخية التي تحتفظ بها		
ث.	التنبؤ بأعداد الطلاب المستجدين المتوقع التحاقهم بكل كلية\معهد للفصل\العام القادم بناء علي الإحصائيات التي تنشرها الجهات الحكومية بأعداد طلاب المدارس الثانوية		
ج.	التنبؤ بأعداد الطلاب المستجدين المتوقع التحاقهم بكل كلية\معهد للفصل\العام القادم بناء علي السجلات التاريخية التي تحتفظ بها بالإضافة الإحصائيات التي تنشرها الحكومة بأعداد طلاب المدارس الثانوية		
ح.	بناء علي البيانات التاريخية التي تحتفظ بها نستطيع تقديم نصيحة للطالب\الطالبة بخصوص إختيار أفضل تخصص يمكنهم الالتحاق به		
خ.	تأجيل قبول الطالب المستجد للفصل\العام القادم		
د.	قبول أو رفض الطالب المحول من كلية\معهد آخر		
ذ.	قبول أو رفض الطالب المحول من كلية\معهد آخر بناء علي السجلات التاريخية التي تحتفظ بها		
•	**قرارات مرتبطة بعملية التسجيل:**		
ر.	التنبؤ بأداء الطالب التعليمي بناء علي السجلات التاريخية التي تحتفظ بها		
ز.	التنبؤ بنتائج الطلاب في مادة معينة بناء علي السجلات التاريخية التي تحتفظ بها		
س.	تقسيم الطلاب إلي مجموعات متشابهة من حيث الإنجاز التعليمي		
ش.	التنبؤ بأداء الطالب التعليمي بناء علي المجموعة التي ينتمي إليها		
ص.	وضع الطالب علي قائمة الإنذار		
ض.	التنبؤ بالطلاب المتوقع ظهورهم في قائمة الإنذار بناء علي السجلات التي تحتفظ بها		
ط.	عمل علاقة بين أداء الطلاب التعليمي و الأقسام التي ينتمون إليها		
ظ.	التنبؤ بأعداد الطلاب المتوقعين في المواد المختلفة		
ع.	إتخاذ القرار بفصل أحد الطلاب		
•	**قرارات أخري تري من الضروري أن يقوم النظام باتخاذها:**		
غ.	.....		
ف.	.....		

ق.	.....
ك.	.....
ل.	.....

٢٤. الجدول التالي يحتوي علي مجموعة من الوظائف المرتبطة بعلمية القبول و التسجيل في الجامعات. من فضلك إختار (نعم) للوظائف التي تري أن نظام معلومات القبول و التسجيل المثالي يجب أن يحتوي عليها، أو (لا) للوظائف التي لا يجب أن يحتوي عليها نظام المعلومات المثالي. أستخدم العلامة (✓) لعمل اختياراتك. تذكر أنه لا توجد إجابات صحيحة و إجابات خاطئة:

رقم	إسم الوظيفة	نعم	لا
أ.	التنبؤ بأداء الطلاب الجدد		
ب.	التنبؤ بأداء الطلاب القدامى		
ت.	طباعة تقارير تصف حالة الطلاب مثال: شهادة بالمعدل \ التقديرات transcripts		
ث.	استخراج إحصائيات مختلفة		
ج.	تقسيم الطلاب إلي مجموعات متشابهة		
ح.	إستخدام البيانات التاريخية		
خ.	إستخدام البيانات الخارجية مثال: التشريعات الحكومية		
د.	ايجاد علاقات بين بيانات الطلاب و بعضها		
ذ.	إستخراج تقارير إستثنائية غير موجودة أصلا في النظام		
ر.	أخري		
ز.	أخري		

٢٥. الجدول التالي يحتوي علي مجموعة من خصائص نظم المعلومات. من فضلك إختار (نعم) للخصائص التي تـري أن نظام المعلومات المثالي يجب أن يحتوي عليها، أو (لا) للخصائص التي لا يجب أن يحتوي عليها نظام المعلومـات المثالي. أستخدم العلامة (✓) لعمل اختياراتك. تذكر أنه لا توجد إجابات صحيحة و إجابات خاطئة:

رقم	إسم الخاصية	نعم	لا
أ.	سهولة الإستخدام		
ب.	يتطلب وقت قليل للتعلم		
ت.	مشاركة المستخدم أثناء بناء النظام		
ث.	قابلية النظام للنمو في المستقبل		
ج.	نظام مرن		
ح.	يمكنه الاندماج مع أنظمة معلومات أخري		
خ.	إمكانية استخدام البريد الإلكتروني من داخل النظام		
د.	إمكانية تشغيل النظام من خلال الإنترنت		
ذ.	انخفاض تكلفة إنشاء النظام		
ر.	أخري		
ز.	أخري		

شكرا علي وقتك و تعاونك معنا ،

من فضلك أرسل هذه النسخة بعد ملئها مستخدما المظروف المرسل

Appendix (B)

Respondents distribution

AUC

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		**Total**
University-level						1					**1**
Faculty of Science											
Faculty of Commerce											
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents						**1**					**1**

Table (B-1). AUC response.

-ROW (S): 112

CITY

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level						1		1			**2**
Faculty of Science											
Faculty of Commerce											
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents						**1**		**1**			**2**

Table (B-2). City response.

-ROW (S): 2, 3

AASTMT

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level						1		3		34	**38**
Faculty of Science											
Faculty of Commerce											
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering						1					**1**
Faculty of Home Economics											
College of Management				1		2					**3**
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering						1					**1**
College of Maritime Studies						1					**1**
DSS Unit											
Total respondents				**1**		**6**		**3**		**34**	**44**

Table (B-3). AASTMT response.

-ROW (S): 64, 65, 76-89, 117, 120, 121, 132, 133, 135, 142, 146, 147, 151-158

MUST

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level						1		1		1	**3**
Faculty of Science											
Faculty of Commerce											
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents						**1**		**1**		**1**	**3**

Table (B-4). MUST response.

-ROW (S): 4, 5, 6

ALEX

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level											
Faculty of Science											
Faculty of Commerce	1				2						**3**
Faculty of Law											
Faculty of Hotels and Tourism	1		1		1						**3**
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts			1						2		**3**
Faculty of Agriculture	1		1		1						**3**
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents	**3**		**3**		**4**				**2**		**12**

Table (B-5). ALEX response.

-ROW (S): 53, 54, 93, 94, 95, 138, 139, 140, 143, 144, 148, 149

ASSIUT

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level											
Faculty of Science	1		1		1						**3**
Faculty of Commerce	1		1		1						**3**
Faculty of Law	1		1		1						**3**
Faculty of Hotels and Tourism											
Faculty of Education	1		1		1						**3**
Faculty of Medicine											
Faculty of Physical Education					1						**1**
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts					1						**1**
Faculty of Agriculture	1		1		1						**3**
Faculty of Engineering	1		1		1						**3**
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services	1		1		1						**3**
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents	**7**		**7**		**9**						**23**

Table (B-6). ASSIUT response.

-ROW (S): 7-29

TANTA

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level											
Faculty of Science	1		1		1				1		**4**
Faculty of Commerce											
Faculty of Law	1				1						**2**
Faculty of Hotels and Tourism											
Faculty of Education			3		2						**5**
Faculty of Medicine			1								**1**
Faculty of Physical Education											
Faculty of Dentist			1		1						**2**
Faculty of Pharmacy	1		1		1						**3**
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents	**3**		**7**		**6**				**1**		**17**

Table (B-7). TANTA response.

-ROW (S): 55-59, 115, 116, 123-131, 145

ZAGAZIG

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level											
Faculty of Science			1		1						**2**
Faculty of Commerce					1						**1**
Faculty of Law					1						**1**
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine					1						**1**
Faculty of Physical Education					1						**1**
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary					1						**1**
Faculty of Arts					1						**1**
Faculty of Agriculture					1						**1**
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents			**1**		**8**						**9**

Table (B-8). ZAGAZIG response.

-ROW (S): 73, 74, 90, 91, 92, 110, 111, 113, 114

MENOUFIA

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level											
Faculty of Science	1		1						1		**3**
Faculty of Commerce			1		1						**2**
Faculty of Law	1				1						**2**
Faculty of Hotels and Tourism											
Faculty of Education	1		1		1						**3**
Faculty of Medicine	1				1						**2**
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts	1		1						1		**3**
Faculty of Agriculture	1		1						1		**3**
Faculty of Engineering	1		1		1						**3**
Faculty of Home Economics	1		1						1		**3**
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents	**8**		**7**		**5**				**4**		**24**

Table (B-9). MENOUFIA response.

-ROW (S): 30-52, 75

SUEZ CANAL

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level											
Faculty of Science	1		1						1		**3**
Faculty of Commerce									1		**1**
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education			2		2						**4**
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy	1		1						1		**3**
Faculty of Veterinary	1										**1**
Faculty of Arts											
Faculty of Agriculture	2										**2**
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics	1										**1**
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit									6		**6**
Total respondents	**6**		**4**		**2**				**9**		**21**

Table (B-10). SUEZ CANAL response.

-ROW (S): 60-63, 96-109, 122, 136, 137

SENGOR

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level								1			**1**
Faculty of Science											
Faculty of Commerce											
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering											
Faculty of Home Economics											
College of Management											
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents								**1**			**1**

Table (B-11). SENGOR response.

-ROW (S): 150

MIU

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level						1		1			**2**
Faculty of Science											
Faculty of Commerce											
Faculty of Law											
Faculty of Hotels and Tourism											
Faculty of Education											
Faculty of Medicine											
Faculty of Physical Education											
Faculty of Dentist											
Faculty of Pharmacy											
Faculty of Veterinary											
Faculty of Arts											
Faculty of Agriculture											
Faculty of Engineering								1		1	
Faculty of Home Economics											
College of Management						4		2			
Faculty of Informatics											
Faculty of Social Services											
College of Marine Engineering											
College of Maritime Studies											
DSS Unit											
Total respondents						**5**		**4**		**1**	**10**

Table (B-12). MIU response.

-ROW (S): 159-168

South Valley

	Dean	Asso. Dean	Registrar	Adm. Officer	Others; Sen. Acd. Advisor or Director	**Total**
University-level						
Faculty of Science						
Faculty of Commerce						
Faculty of Law						
Faculty of Hotels and Tourism						
Faculty of Education						
Faculty of Medicine						
Faculty of Physical Education						
Faculty of Dentist						
Faculty of Pharmacy						
Faculty of Veterinary						
Faculty of Arts						
Faculty of Agriculture						
Faculty of Engineering						
Faculty of Home Economics						
College of Management						
Faculty of Informatics						
Faculty of Social Services						
College of Marine Engineering						
College of Maritime Studies						
DSS Unit						
Total respondents						

Questionnaires returned from respondents but lost in mail, where other trails to re-fill them again failed!

Table (B-13). SOUTH VALLEY response.

-ROW (S): -

Appendix (C)

Codes

Codes of the 2000 records[1]

Nationalities		
	0	Unknown
	1	Jordan
	2	Algeria
	3	Sudan
	4	Syria
	5	Somalia
	6	Iraq
	7	Palestine
	8	Qatar
	9	Lebanon
	10	Libya
	11	Egypt
	12	Yemen
	13	Kuwait
	14	Ethiopia
	15	Tanzania
	16	Gambia
	17	Moon Lands
	18	Sierra Lion
	19	Ghana
	20	Cameroon
	21	Liberia
	22	Namibia
	23	Nigeria
	24	Turkey
	25	Saudi
	26	Madaghashkar
	27	Gabon
	28	Eritrea
	29	Oman
	30	Kenya
	31	Italy
	32	Emirates
	33	Bahrain
	34	Australia

[1] The researcher did not create this coding system, it has been used by the AASTMT.

	35	Pakistan
	36	Cyprus
	37	Mauritania
	38	Terkistan
	39	Indonesia
	40	Morocco
	41	India
	42	South Africa
	43	Tunisia
	44	USA
	45	Canada
	46	UK
	47	Gheynia
	48	Sweden
	49	Germany
	50	Singapore
	51	Senegal
	52	Iran
	53	Russia
	54	Not-Kuwait
	55	Djibouti
	56	Zambia
	57	Afghanistan
	58	Croatia
	59	Dominican
	60	India
	61	Poland
	62	Greece
	63	Holland
	214	Philippine
	222	Bangladesh
Batches		
2000	1	151
1999	3	150
1999	2	149
1999	1	137
1998	3	136
1998	2	135

1998	1	134
1997	3	133
1997	2	132
1997	1	127
1996	3	126
1996	2	105
1996	1	104
1995	3	103
1995	2	102
1995	1	101
1994	4	100
1994	3	99
1994	2	98
1994	1	97
1993	4	96
1993	3	95
1993	2	94
1993	1	93
1992	4	92
1992	3	91
1992	2	90
1992	1	89
1991	4	88
1991	3	87
1991	2	86
1991	1	85
1990	4	84
1990	3	83
1990	2	82
1990	1	81
High Schools		
	1	Thanwya Amma- Math
	2	Thanwya Amma- Science
	3	Thanwya Azhar
	4	Preparatory Diploma
	5	Thanwya Amma- Arts
	6	IGCSE
	7	Swedish
	8	Else

	100	B.Sc. Comm-Eng.
	101	Maritime Transp. Diploma
	102	B.Sc. Law
	103	B.Sc. Maritime Transp.
	104	Public Law Diploma
	105	Private Law Diploma
	106	B.Sc. Eye Medicine
	107	B.Sc. Hot. And Tourism
	108	Nursing Diploma
	111	Thanwya Amma- New
	112	American Diploma- Science
	113	English Certi. - Science
	114	Courses Thanwya- Science
	115	Thanwya Azhar- Science
	116	Thanwya -Commerce
	117	Else- Science
	118	Zanzibar Cert.
	121	Thanwya Amma New- Science
	122	American Diploma- Arts
	123	English Certi. - Science
	124	Courses Thanwya- Arts
	125	Thanwya Azhar- Arts
	126	Technical Thanwya- Industry
	127	Else- Arts
	131	Thanwya Amma New- Arts
	133	IGCSE- Old
	135	Thanwya Azhar- Old
	137	Swedish
	139	Preparatory- Maritime Transp.
	141	Thanwya Amma Old- Science
	147	Diploma- Preparatory
	151	Thanwya Amma Old- Arts
	157	Else
	161	Thanwya Amma Old- Science
	171	Thanwya Amma Old- Math
	211	Thanwya Amma New- Lang. Sch.
	214	Courses Thanwya Science- Lang.
	217	Else- Science- Lang.
	221	Thanwya Amma New- Lang. Sch. Science
	224	Courses Thanwya Arts- Lang.

	227	Else- Arts- Lang
	231	Thanwya Amma New- Lang. Sch. Arts
	241	Thanwya Amma Old- Lang. Sch. Science
	251	Thanwya Amma Old- Lang. Sch. Arts
	501	3rd Officer
	502	B.Sc. Maritime
	503	B.Sc. Marine Sciences
	950	B.Sc. Arts
	951	B.Sc. Engineering
	952	B.Sc. Comm.
	953	B.Sc. Science
	954	German Diploma
	955	B.Sc. Fine Arts
	956	B.Sc. Agriculture
	957	High Diploma
	958	Computer Diploma
	959	1st Mate
	960	2nd Officer
	961	Maritime- Basic Studies
	9001	B.Sc.
	9004	Le' cans
	9040	B.Sc. Hotels
	9041	B.Sc. Computing
	9053	B.Sc. law
	9054	B.Sc. Social Service
	9055	B.Sc. Cooperation Inst.
	9059	B.Sc. Maritime Science
	9060	B.Sc. Military Science
	9066	Fellow, School of war
	9077	B.Sc. Physical Educ.
	9078	B.Sc. Dentistry
	9088	B.Sc. Comm. 2
	9089	B.Sc. Medicine
	9092	B.Sc. Education
	9093	Coach- swimming
	9094	2nd officer
	9100	Basic theoretical studies
	9112	Maritime Eng.

Gender		
	1	Male
	0	Female
Grades		
	1	Pass
	2	Good
	3	Very Good
	4	Excellent
	5	V. Good - Honor
	6	Excellent - Honor
	0	Poor
Degrees		
	1	BBA English section
	2	BBA Arabic section
	3	Bachelor of Maritime Transport
	4	BTech. Electronics
	5	BTech. Marine Eng.
	6	Bachelor of Hotels and Tourism
	7	Bachelor of Maritime
	8	B.Sc. Computers
	9	B.Sc. Electronics
	10	B.Sc. Marine Eng.
	11	B.Sc. Mechanical Eng.

Table (C-1). Data records' codes.

Appendix (D)

Data Warehouse Design

1. DW Source-to-target fields' matrix

The Admission and Registration functions DW depends on the operational systems. The next matrices show how the operational system DB fields have been mapped into the DW fields.

			No	1										
			Schema											
			Table	**Dimension: COLLEGE**										
No	System	Table	Field											
1	**ODS**	**COLLEGE**	COL_NAME	X										
			COL_SERIAL **PK**[1]		X									
			COL_LOCATION			X								
2	**ODS**	**DEPARTMENT**	DEP_TITLE	X										
			DEP_ID **PK**		X									
			DEP_LOCATION			X								
			DEP_TYPE				X							
3	**ODS**	**MAJOR**	MAJ_TITLE	X										
			MAJ_SNO **PK**		X									
			MAJ_MIN_HIGH_SCH_PERCENT			X								
			No	2										
			Schema											
			Table	**Dimension: APPLICANT**										
No	System	Table	Field											
4	**ODS**	**APPLICANT**	/APP_AGE	X										
			/APP_PRED_GRAD_GRADE_M		X									
			/APP_PRED_MAJOR_F			X								
			APP_FULL_NAME				X							
			APP_GENDER					X						
			APP_TELEPHONE						-					
			APP_CERT_PERCENTAGE							X				
			APP_ADDRESS								X			
			APP_PREVIOUSLY_ABANDONED									X		
			APP_TRANSFERRED										X	.,[2]
			APP_ INTERVIEWED											X
			APP_DOB											X
			APP_CODE **PK**											X
			APP_SEC_CERT_YEAR											X
			/APP_HOLD_NEXT_BATCH_G											X
			/APP_ACCEPT_A											X
5	**ODS**	**NATIONALITY**	NAT_IDENTIFICATION **PK**	X										
			NAT_NATIONALITY		X									
6	**ODS**	**CERTIFICATE**	CER_NUMBER **PK**	X										
			CER_NAME		X									
			SEC_ORIGIN			X								
7	**ODS**	**UNIVERSITY**	UNI-NAME	X										
			UNI_ID **PK**		X									
			UNI_COUNTRY			X								
			No	3										
			Schema											
			Table	**Dimension: SEMESTER**										
8	**ODS**	**BATCH**	BAT_CEILING	X										
			BAT_NUMBER **PK**		X									
			BAT_TITLE			X								
			/BAT_PRED_APPLICANTS_C				X							
			BAT_YEAR					-						

[1] PK stands for the primary key of the entity in the OLTP system, but not the PK of neither the FACT not the DIMENSION table.
[2] Due to space constraint, this column could be used to represent more than one field.

			BAT_OPEN_DATE							X				
			BAT_CLOSING_DATE								X			
			BAT_MARKET_SHARE									X		
			BAT_GOVERNMENT_STATISTICS										X	
9	**ODS**	**SEMESTER**	SEM_CODE **PK**	X										
			SEM_NAME		X									
			SEM_YEAR			X								
			No	**4**										
			Schema											
			Table	**Dimension: COURSE**										
No	System	Table	Field											
10	**ODS**	**COURSE**	COU_TITLE	X										
			COU_CODE **PK**		X									
			COU_CREDIT_HOURS			X								
			COU_PASS_MARK				X							
			COU_FULL_MARK					X						
			COU_STAGE						X					
			COU_AREA							X				
11	**ODS**	**PREREQUISITE**	PRE_SNO **PK**	X										
			PRE_DETAIL1		X									
			PRE_DETAIL2			X								
			PRE_DETAIL3				X							
			PRE_DETAIL4					X						
12	**ODS**	**LAB**	LAB_TITLE	X										
			LAB_DESCRIBTION		-									
			LAB_CODE **PK**			X								
13	**ODS**	**COURSE_ MAJOR**	COU_MAJ_COUNTER **PK**	X										
			No	**5**										
			Schema											
			Table	**Dimension: ASSISTANTSHIP**										
No	System	Table	Field											
14	**ODS**	**ASSISTANTSHIP**	ASS_TITLE	X										
			ASS_NUMBER **PK**		X									
			ASS_CATEGORY			X								
			ASS_REQUIREMENTS				X							
			ASS_CATEGORY					X						
			ASS_DISCOUNT_RATE						X					
			No	**6**										
			Schema											
			Table	**Dimension: PENALTY**										
15	**ODS**	**PENALTY**	PEN_NAME	X										
			PEN_SERIAL **PK**		X									
			PEN_CONCEQUENCES			X								
			No	**7**										
			Schema											
			Table	**Dimension: STUDENT**										
No	System	Table	Field											
16	**ODS**	**STUDENT**	STU_REGISTRATION_NO **PK**	X										
			/STU_TOTAL_COURSES		X									
			/STU_PREDICT_PERFORMANCE_J			X								
			/STU_GPA				X							
			/STU_TOT_CREDIT_HOURS_REG					X						
			/STU_TOT_CREDIT_HOURS_ACH						X					
			/STU_GRADUATION_STATE							X				
			/STU_ABANDONMENT_STATE_R								X			
			/STU_ON_PROBATION_STATE_N									X		

			/STU_PREDICT_ON_PROBATIO_O										X	
17	**ODS**	**AUTHORITY**	AUT_NAME AUT_ADDRESS AUT_TELEPHONE AUT_CONTACT_PERSON AUT_CODE **PK**	X	-	-	-	X						
18	**ODS**	**STUDENT_ ASSISTANTSHIP**	STU_ASS_STARTS STU_ASS_ENDS STU_ASS_SEIAL_NO **PK**	X	X	X								
19	**ODS**	**PENALTY_ STUDENT**	PEN_STU_SNO **PK** PEN_STU_DATE	X	X									
20	**ODS**	**GPA**	GPA_COUNTER **PK**	X										
			No	**8**										
			Schema											
			Table	**Dimension: TUITION**										
No	System	Table	Field											
21	**ODS**	**TUITION**	TUI_COUNTER **PK** TUI_AMOUNT TUI_CURRENCY	X	X	X								
			No	**9**										
			Schema											
			Table	**Dimension: PAYMENT**										
No	System	Table	Field											
22	**ODS**	**PAYMENT**	/PAY_AMOUNT PAY_DATE /PAY_NET_AMOUNT PAY_DISCOUNTED /PAY_DISCOUNT PAY_CURRENCY PAY_METHOD PAY_RECEIPT_NO **PK**	X	X	X	X	X	X	X	X			
			No	**10**										
			Schema											
			Table	**Dimension: REGISTRATION**										
No	System	Table	Field											
23	**ODS**	**REGISTRATION**	REG_SERIALNO **PK** REG_DATE	X	X									
			No	**11**										
			Schema											
			Table	**Dimension: EXAM**										
No	System	Table	Field											
24	**ODS**	**EXAM**	EXA_NUMBER **PK** EXA_TYPE	X	X									
			No	**12**										
			Schema											
			Table	**Dimension: MARK**										
No	System	Table	Field											
25	**ODS**	**MARK**	MAR_RES_COUNTER **PK** MAR_RES_DATE MAR_RES_MARK /MAR_POINTS	X	X	X	X							
			No	**13**										
			Schema											
			Table	**Fact: STUDENT_RECORD**										

No	System	Table	Field											
-			**COLLEGE KEY**	X										
			APPLICANT KEY		X									
			COURSE KEY			X								
			ASSISTANTSHIP KEY				X							
			PENALTY KEY					X						
			STUDENT KEY						X					
			TUITION KEY							X				
			PAYMENT KEY								X			
			REGISTRATION KEY									X		
			EXAM KEY										X	
			MARK KEY											X
		-	**SEMESTER KEY**											X
			STU_REC_AVERAGE_GPA											X
			STU_REC_SUM_PAYMENTS											X
			STU_REC_SUM_YEAR_IN_UNIVERS											X
			STU_REC_AVERAGE_DISCOUNTS											X
			STU_REC_COUNT_COURSES_PASS											X
			STU_REC_COUNT_COURSES_FAIL											X
			STU_REC_COUNT_ALL_COURSES											X
			STU_REC_COUNT_PENALTIES											X
			STU_REC_COUNT_ASSISTANTS											X
			STU_REC_COUNT_MAJORS											X
			STU_REC_COUNT_MARKS											

Table (D-1). Source-to-target field matrix.

2. DW CREATE Statements

```
/* Microsoft SQL Server - Scripting                */
/* Server: NTSERVER                                */
/* Database: ARDSS_DW                              */
/* Creation Date 7/23/01 5:29:43 PM                      */

set quoted_identifier on
GO

/****** Object:  Login aelragal    Script Date: 7/23/01 5:29:44 PM ******/
if not exists (select * from master..syslogins where name = 'aelragal')
BEGIN
      declare @logindb varchar(30), @loginlang varchar(30) select @logindb =
'ARDSS_DB', @loginlang = null
      if @logindb is null or not exists (select * from master..sysdatabases where
name = @logindb)
            select @logindb = 'master'
      if @loginlang is null or (not exists (select * from master..syslanguages
where name = @loginlang) and @loginlang <> 'us_english')
            select @loginlang = @@language
      exec sp_addlogin 'aelragal', null, @logindb, @loginlang
END
GO

/****** Object:  User aelragal    Script Date: 7/23/01 5:29:44 PM ******/
if not exists (select * from sysusers where name = 'aelragal' and uid < 16382)
      EXEC sp_adduser 'aelragal', 'aelragal', 'public'
GO

/****** Object:  Table dbo.student_record    Script Date: 7/23/01 5:29:45 PM
******/
if exists (select * from sysobjects where id = object_id('dbo.student_record') and
sysstat & 0xf = 3)
      drop table "dbo"."student_record"
GO

/****** Object:  Table dbo.applicant    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.applicant') and
sysstat & 0xf = 3)
      drop table "dbo"."applicant"
GO

/****** Object:  Table dbo.assistantship    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.assistantship') and
sysstat & 0xf = 3)
      drop table "dbo"."assistantship"
GO

/****** Object:  Table dbo.college    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.college') and sysstat
& 0xf = 3)
      drop table "dbo"."college"
GO

/****** Object:  Table dbo.course    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.course') and sysstat
& 0xf = 3)
      drop table "dbo"."course"
GO
```

```
/****** Object:  Table dbo.exam    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.exam') and sysstat &
0xf = 3)
      drop table "dbo"."exam"
GO

/****** Object:  Table dbo.mark    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.mark') and sysstat &
0xf = 3)
      drop table "dbo"."mark"
GO

/****** Object:  Table dbo.payment    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.payment') and sysstat
& 0xf = 3)
      drop table "dbo"."payment"
GO

/****** Object:  Table dbo.penalty    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.penalty') and sysstat
& 0xf = 3)
      drop table "dbo"."penalty"
GO

/****** Object:  Table dbo.registration    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.registration') and
sysstat & 0xf = 3)
      drop table "dbo"."registration"
GO

/****** Object:  Table dbo.semester    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.semester') and
sysstat & 0xf = 3)
      drop table "dbo"."semester"
GO

/****** Object:  Table dbo.student    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.student') and sysstat
& 0xf = 3)
      drop table "dbo"."student"
GO

/****** Object:  Table dbo.tuition    Script Date: 7/23/01 5:29:45 PM ******/
if exists (select * from sysobjects where id = object_id('dbo.tuition') and sysstat
& 0xf = 3)
      drop table "dbo"."tuition"
GO

/****** Object:  Table dbo.applicant    Script Date: 7/23/01 5:29:45 PM ******/
CREATE TABLE "dbo"."applicant" (
      "applicant_key" numeric(8, 0) NOT NULL ,
      "app_age" numeric(3, 0) NULL ,
      "app_predicted_graduation_grade" char (20) NULL ,
      "app_predicted_major_f" char (30) NULL ,
      "app_full_name" char (35) NULL ,
      "app_dob" "datetime" NULL ,
      "app_gender" char (1) NULL ,
      "app_address" char (35) NULL ,
      "app_previously_abandoned" char (1) NULL ,
      "app_transferred" char (1) NULL ,
```

```
	"app_code" numeric(6, 0) NULL ,
	"app_certificate_year" numeric(5, 0) NULL ,
	"app_certification_percent" numeric(4, 0) NULL ,
	"app_interviewed" char (1) NULL ,
	"app_accept" char (1) NULL ,
	"app_hold_to_next_batch_g" char (1) NULL ,
	"uni_name" char (35) NULL ,
	"uni_id" numeric(5, 0) NULL ,
	"uni_country" char (15) NULL ,
	"nat_identification" numeric(5, 0) NULL ,
	"nat_nationality" char (25) NULL ,
	"cer_number" numeric(5, 0) NULL ,
	"cer_name" char (35) NULL ,
	"cer_origin" char (25) NULL ,
	"bat_number" numeric(5, 0) NULL ,
	CONSTRAINT "PK___10__13" PRIMARY KEY  CLUSTERED
	(
		"applicant_key"
	)
)
GO

 CREATE  INDEX "app_age_ix" ON "dbo"."applicant"("app_age") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "app_cert_percent_ix" ON
"dbo"."applicant"("app_certification_percent") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "app_cert_yr_ix" ON "dbo"."applicant"("app_certificate_year") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "app_certificate_year_ix" ON
"dbo"."applicant"("app_certificate_year") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "app_code_ix" ON "dbo"."applicant"("app_code") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "cer_number_ix" ON "dbo"."applicant"("cer_number") WITH  FILLFACTOR
= 75
GO

 CREATE  INDEX "nat_identification_ix" ON "dbo"."applicant"("nat_identification")
WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "uni_id_ix" ON "dbo"."applicant"("uni_id") WITH  FILLFACTOR = 75
GO

/****** Object:  Table dbo.assistantship    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."assistantship" (
	"assistantship_key" numeric(8, 0) NOT NULL ,
	"ass_title" "text" NULL ,
	"ass_number" numeric(5, 0) NOT NULL ,
	"ass_requirements" "text" NULL ,
	"ass_discount_rate" numeric(4, 0) NULL ,
	"ass_category" "text" NULL ,
	CONSTRAINT "PK___7__13" PRIMARY KEY  CLUSTERED
	(
```

```
            "assistantship_key"
      )
)
GO

 CREATE  INDEX "ass_discount_rate_ix" ON "dbo"."assistantship"("ass_discount_rate")
WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "ass_number_ix" ON "dbo"."assistantship"("ass_number") WITH
FILLFACTOR = 75
GO

/****** Object:  Table dbo.college    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."college" (
      "college_key" numeric(8, 0) NOT NULL ,
      "col_name" "text" NULL ,
      "col_serial" numeric(5, 0) NOT NULL ,
      "col_location" "text" NULL ,
      "dep_title" "text" NULL ,
      "dep_id" numeric(5, 0) NOT NULL ,
      "dep_location" "text" NULL ,
      "dep_type" "text" NULL ,
      "maj_title" char (35) NULL ,
      "maj_sno" numeric(5, 0) NULL ,
      "maj_min_high_school_percent" numeric(5, 1) NULL ,
      CONSTRAINT "PK___8__13" PRIMARY KEY  CLUSTERED
      (
            "college_key"
      )
)
GO

 CREATE  INDEX "col_serial_ix" ON "dbo"."college"("col_serial") WITH  FILLFACTOR =
75
GO

 CREATE  INDEX "dep_id_ix" ON "dbo"."college"("dep_id") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "maj_min_high_sch_perc_ix" ON
"dbo"."college"("maj_min_high_school_percent") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "maj_sno_ix" ON "dbo"."college"("maj_sno") WITH  FILLFACTOR = 75
GO

/****** Object:  Table dbo.course    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."course" (
      "course_key" numeric(8, 0) NOT NULL ,
      "cou_stage" "text" NULL ,
      "cou_title" "text" NULL ,
      "cou_code" "text" NOT NULL ,
      "cou_credit" numeric(3, 0) NULL ,
      "cou_pass_mark" numeric(5, 0) NULL ,
      "cou_full_mark" numeric(5, 0) NULL ,
      "cou_area" "text" NULL ,
      "cou_maj_counter" numeric(5, 0) NULL ,
      "lab_code" numeric(5, 0) NULL ,
      "lab_title" "text" NULL ,
      "pre_sno" numeric(5, 0) NULL ,
```

```
	"pre_detail1" "text" NULL ,
	"pre_detail2" "text" NULL ,
	"pre_detail3" "text" NULL ,
	"pre_detail4" "text" NULL ,
	"pre_detail5" "text" NULL ,
	CONSTRAINT "PK___9__13" PRIMARY KEY  CLUSTERED
	(
		"course_key"
	)
)
GO

 CREATE  INDEX "cou_credit_ix" ON "dbo"."course"("cou_credit") WITH  FILLFACTOR =
75
GO

 CREATE  INDEX "cou_full_mark_ix" ON "dbo"."course"("cou_full_mark") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "cou_maj_counter_ix" ON "dbo"."course"("cou_maj_counter") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "cou_pass_mark_ix" ON "dbo"."course"("cou_pass_mark") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "lab_code_ix" ON "dbo"."course"("lab_code") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "pre_sno_ix" ON "dbo"."course"("pre_sno") WITH  FILLFACTOR = 75
GO

/****** Object:  Table dbo.exam    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."exam" (
	"exam_key" numeric(8, 0) NOT NULL ,
	"exa_number" numeric(5, 0) NOT NULL ,
	"exa_type" "text" NULL ,
	"exa_fk_cou_code" char (8) NULL ,
	CONSTRAINT "PK___5__13" PRIMARY KEY  CLUSTERED
	(
		"exam_key"
	)
)
GO

 CREATE  INDEX "exa_fk_cou_code_ix" ON "dbo"."exam"("exa_fk_cou_code") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "exa_number_ix" ON "dbo"."exam"("exa_number") WITH  FILLFACTOR = 75
GO

/****** Object:  Table dbo.mark    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."mark" (
	"mark_key" numeric(8, 0) NOT NULL ,
	"mar_counter" numeric(6, 0) NOT NULL ,
	"mar_date" "datetime" NULL ,
	"mar_points" "int" NULL ,
	"mar_stu_reg_no" "float" NULL ,
```

```
	"mar_exa_exa_no" "int" NULL ,
	CONSTRAINT "PK___4__13" PRIMARY KEY  CLUSTERED
	(
		"mark_key"
	)
)
GO

 CREATE  INDEX "mar_counter_ix" ON "dbo"."mark"("mar_counter") WITH  FILLFACTOR =
75
GO

 CREATE  INDEX "mar_exa_exa_no_ix" ON "dbo"."mark"("mar_stu_reg_no") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "mar_points_ix" ON "dbo"."mark"("mar_points") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "mar_stu_reg_no_ix" ON "dbo"."mark"("mar_stu_reg_no") WITH
FILLFACTOR = 75
GO

/****** Object:  Table dbo.payment    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."payment" (
	"payment_key" numeric(8, 0) NOT NULL ,
	"pay_receipt_no" numeric(10, 0) NOT NULL ,
	"pay_date" "datetime" NULL ,
	"pay_method" "text" NULL ,
	"pay_amount" numeric(7, 0) NULL ,
	"pay_currency" "text" NULL ,
	"pay_discounted" "text" NULL ,
	"pay_discount" "text" NULL ,
	"pay_net_amount" numeric(6, 0) NULL ,
	"pay_fk_stu_reg_no" "float" NULL ,
	"pay_fk_tui_tui_counter" "int" NULL ,
	CONSTRAINT "PK___6__13" PRIMARY KEY  CLUSTERED
	(
		"payment_key"
	)
)
GO

 CREATE  INDEX "pay_amount_ix" ON "dbo"."payment"("pay_amount") WITH  FILLFACTOR =
75
GO

 CREATE  INDEX "pay_fk_stu_reg_no_ix" ON "dbo"."payment"("pay_fk_stu_reg_no") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "pay_fk_tui_tui_counter_ix" ON
"dbo"."payment"("pay_fk_tui_tui_counter") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "pay_net_amount_ix" ON "dbo"."payment"("pay_net_amount") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "pay_receipt_no" ON "dbo"."payment"("pay_receipt_no") WITH
FILLFACTOR = 75
```

```
GO

/****** Object:  Table dbo.penalty    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."penalty" (
      "penalty_key" numeric(8, 0) NOT NULL ,
      "pen_name" "text" NULL ,
      "pen_serial" numeric(5, 0) NOT NULL ,
      "pen_concequences" "text" NULL ,
      CONSTRAINT "PK___1__13" PRIMARY KEY  CLUSTERED
      (
            "penalty_key"
      )
)
GO

 CREATE  INDEX "pen_serial_ix" ON "dbo"."penalty"("pen_serial") WITH  FILLFACTOR =
75
GO

/****** Object:  Table dbo.registration    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."registration" (
      "registration_key" numeric(8, 0) NOT NULL ,
      "reg_serialNo" numeric(6, 0) NOT NULL ,
      "reg_date" "datetime" NULL ,
      "reg_fk_cou_cou_code" char (8) NULL ,
      "reg_fk_stu_stu_reg_no" "float" NULL ,
      "reg_fk_pay_pay_receipt_no" "float" NULL ,
      CONSTRAINT "PK___2__13" PRIMARY KEY  CLUSTERED
      (
            "registration_key"
      )
)
GO

 CREATE  INDEX "reg_fk_pay_pay_receipt_no" ON
"dbo"."registration"("reg_fk_pay_pay_receipt_no") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "reg_fk_stu_stu_reg_ix" ON
"dbo"."registration"("reg_fk_stu_stu_reg_no") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "reg_serial_no_ix" ON "dbo"."registration"("reg_serialNo") WITH
FILLFACTOR = 75
GO

/****** Object:  Table dbo.semester    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."semester" (
      "semester_key" numeric(8, 0) NOT NULL ,
      "sem_code" numeric(5, 0) NULL ,
      "sem_name" char (12) NULL ,
      "sem_year" numeric(5, 0) NULL ,
      "bat_ceiling" numeric(5, 0) NULL ,
      "bat_number" numeric(5, 0) NULL ,
      "bat_title" char (10) NULL ,
      "bat_open_date" "datetime" NULL ,
      "bat_closing_date" "datetime" NULL ,
      "bat_market_share" numeric(5, 0) NULL ,
      "bat_government_statistics" numeric(6, 0) NULL ,
      CONSTRAINT "PK___2__11" PRIMARY KEY  CLUSTERED
      (
```

```
            "semester_key"
      )
)
GO

 CREATE  INDEX "bat_ceiling_ix" ON "dbo"."semester"("bat_ceiling") WITH  FILLFACTOR
= 75
GO

 CREATE  INDEX "bat_government_stat_ix" ON
"dbo"."semester"("bat_government_statistics") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "bat_market_share_ix" ON "dbo"."semester"("bat_market_share") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "bat_number_ix" ON "dbo"."semester"("bat_number") WITH  FILLFACTOR =
75
GO

 CREATE  INDEX "sem_code_ix" ON "dbo"."semester"("sem_code") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "sem_year_ix" ON "dbo"."semester"("sem_year") WITH  FILLFACTOR = 75
GO

/****** Object:  Table dbo.student    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."student" (
      "student_key" numeric(8, 0) NOT NULL ,
      "stu_registration_no" numeric(10, 0) NOT NULL ,
      "stu_total_courses" numeric(2, 0) NULL ,
      "stu_predicted_performance_j" char (20) NULL ,
      "stu_gpa" numeric(4, 0) NULL ,
      "stu_total_credit_hours_reg" numeric(4, 0) NULL ,
      "stu_total_credit_hours_ach" numeric(4, 0) NULL ,
      "stu_graduation_state" char (1) NULL ,
      "stu_abandonment_state_r" char (1) NULL ,
      "stu_on_probation_state_n" char (1) NULL ,
      "stu_predicted_on_probation_sta" char (1) NULL ,
      "stu_ass_serial_no" numeric(6, 0) NULL ,
      "stu_ass_starts" "datetime" NULL ,
      "stu_ass_ends" "datetime" NULL ,
      "aut_name" char (30) NULL ,
      "aut_code" numeric(5, 0) NULL ,
      "gpa_counter" numeric(6, 0) NULL ,
      "pen_stu_sno" numeric(5, 0) NULL ,
      "pen_stu_date" "datetime" NULL ,
      "stu_fk_app_app_code" "int" NULL ,
      CONSTRAINT "PK__11__13" PRIMARY KEY  CLUSTERED
      (
            "student_key"
      )
)
GO

 CREATE  INDEX "aut_code_ix" ON "dbo"."student"("aut_code") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "gpa_counter_ix" ON "dbo"."student"("gpa_counter") WITH  FILLFACTOR
= 75
```

```
GO

 CREATE  INDEX "pen_stu_sno_ix" ON "dbo"."student"("pen_stu_sno") WITH  FILLFACTOR
= 75
GO

 CREATE  INDEX "stu_ass_serial_no_ix" ON "dbo"."student"("stu_ass_serial_no") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "stu_fk_app_app_co_ix" ON "dbo"."student"("stu_fk_app_app_code")
WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "stu_gpa_ix" ON "dbo"."student"("stu_gpa") WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "stu_registration_no_ix" ON "dbo"."student"("stu_registration_no")
WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "stu_total_courses_ix" ON "dbo"."student"("stu_total_courses") WITH
FILLFACTOR = 75
GO

 CREATE  INDEX "stu_total_credit_hours" ON
"dbo"."student"("stu_total_credit_hours_reg") WITH  FILLFACTOR = 75
GO

/****** Object:  Table dbo.tuition    Script Date: 7/23/01 5:29:46 PM ******/
CREATE TABLE "dbo"."tuition" (
      "tuition_key" numeric(8, 0) NOT NULL ,
      "tui_counter" numeric(6, 0) NOT NULL ,
      "tui_amount" numeric(7, 0) NULL ,
      "tui_currency" "text" NULL ,
      "tui_fk_sem_sem_code" "int" NULL ,
      "tui_fk_maj_maj_sno" "int" NULL ,
      CONSTRAINT "PK___3__13" PRIMARY KEY  CLUSTERED
      (
            "tuition_key"
      )
)
GO

 CREATE  INDEX "tui_amount_ix" ON "dbo"."tuition"("tui_amount") WITH  FILLFACTOR =
75
GO

 CREATE  INDEX "tui_counter_ix" ON "dbo"."tuition"("tui_counter") WITH  FILLFACTOR
= 75
GO

 CREATE  INDEX "tui_fk_maj_maj_sno_ix" ON "dbo"."tuition"("tui_fk_maj_maj_sno")
WITH  FILLFACTOR = 75
GO

 CREATE  INDEX "tui_fk_sem_sem_co_ix" ON "dbo"."tuition"("tui_fk_sem_sem_code")
WITH  FILLFACTOR = 75
GO
```

```
/****** Object:  Table dbo.student_record    Script Date: 7/23/01 5:29:46 PM
******/
CREATE TABLE "dbo"."student_record" (
      "college_key" numeric(8, 0) NULL ,
      "applicant_key" numeric(8, 0) NULL ,
      "course_key" numeric(8, 0) NULL ,
      "assistantship_key" numeric(8, 0) NULL ,
      "penalty_key" numeric(8, 0) NULL ,
      "student_key" numeric(8, 0) NULL ,
      "tuition_key" numeric(8, 0) NULL ,
      "payment_key" numeric(8, 0) NULL ,
      "registration_key" numeric(8, 0) NULL ,
      "exam_key" numeric(8, 0) NULL ,
      "mark_key" numeric(8, 0) NULL ,
      "stu_rec_average_gpa" numeric(6, 1) NULL ,
      "stu_rec_sum_payments" numeric(6, 0) NULL ,
      "stu_rec_sum_year_in_university" numeric(5, 0) NULL ,
      "stu_rec_average_discounts" numeric(6, 1) NULL ,
      "stu_rec_count_course_pass" numeric(3, 0) NULL ,
      "stu_rec_count_course_fail" numeric(3, 0) NULL ,
      "stu_rec_count_course_all" numeric(3, 0) NULL ,
      "stu_rec_count_penalty" numeric(3, 0) NULL ,
      "stu_rec_count_assistantship" numeric(3, 0) NULL ,
      "stu_rec_count_majors" numeric(3, 0) NULL ,
      "stu_rec_count_marks" numeric(3, 0) NULL ,
      "student_record" numeric(8, 0) NOT NULL ,
      "semester_key" numeric(8, 0) NULL ,
      CONSTRAINT "PK___12__13" PRIMARY KEY  CLUSTERED
      (
            "student_record"
      ),
      CONSTRAINT "FK___1__13" FOREIGN KEY
      (
            "college_key"
      ) REFERENCES "dbo"."college" (
            "college_key"
      ),
      CONSTRAINT "FK___10__13" FOREIGN KEY
      (
            "exam_key"
      ) REFERENCES "dbo"."exam" (
            "exam_key"
      ),
      CONSTRAINT "FK___11__13" FOREIGN KEY
      (
            "mark_key"
      ) REFERENCES "dbo"."mark" (
            "mark_key"
      ),
      CONSTRAINT "FK___2__13" FOREIGN KEY
      (
            "applicant_key"
      ) REFERENCES "dbo"."applicant" (
            "applicant_key"
      ),
      CONSTRAINT "FK___3__13" FOREIGN KEY
      (
            "course_key"
      ) REFERENCES "dbo"."course" (
            "course_key"
      ),
```

```
	CONSTRAINT "FK___4__13" FOREIGN KEY
	(
		"assistantship_key"
	) REFERENCES "dbo"."assistantship" (
		"assistantship_key"
	),
	CONSTRAINT "FK___5__13" FOREIGN KEY
	(
		"penalty_key"
	) REFERENCES "dbo"."penalty" (
		"penalty_key"
	),
	CONSTRAINT "FK___6__13" FOREIGN KEY
	(
		"student_key"
	) REFERENCES "dbo"."student" (
		"student_key"
	),
	CONSTRAINT "FK___7__13" FOREIGN KEY
	(
		"tuition_key"
	) REFERENCES "dbo"."tuition" (
		"tuition_key"
	),
	CONSTRAINT "FK___8__13" FOREIGN KEY
	(
		"payment_key"
	) REFERENCES "dbo"."payment" (
		"payment_key"
	),
	CONSTRAINT "FK___9__13" FOREIGN KEY
	(
		"registration_key"
	) REFERENCES "dbo"."registration" (
		"registration_key"
	),
	CONSTRAINT "FK_student_record_1__11" FOREIGN KEY
	(
		"semester_key"
	) REFERENCES "dbo"."semester" (
		"semester_key"
	)
)
GO
```

3. The Data Warehouse Reports

The Admission and Registration Managers have requested some reports for different decision situations. The reports have been generated using Crystal Reports[3] (4.5) by Seagate. The names of those reports are:

1. Nationality, Majors, and GPA;
2. Years in University, Majors, and GPA;
3. Majors versus Gender;
4. Age, GPA, and Majors;
5. High Schools, Majors, and GPA;
6. Majors value-added;
7. University value-added;
8. Demand Curve;
9. Gender Distribution;
10. Major Distribution;
11. Applicants' High School Scores, Majors, and Average Graduation Scores.

Following are copies of those reports.

[3] A client release has been used.

- **Nationality, Majors, and GPA**

User (s):

Dean	✓
Deputy Dean	✓
Registrar	
Admission Officer	✓
Other	✓

NATIONALITY, MAJOR, AND AVERAGE GRADUATION GRADES

9/5/01

MAJORS

NATIONALITY	1	2	3	4	5	6	7	8	9	10	11	AVERAGE
1	82.5	61.1	67.5	64.2	74.1		83.7	65.5	70.9	72.7		70.5
3	85.0	67.3	45.0	62.3	59.0	65.0	72.5	61.8	65.3	67.7	68.7	65.8
4	80.0	65.0		60.0	75.0		89.1	68.1	65.0	67.5		71.3
6		50.0							72.5	80.0		72.5
7	57.5	59.1	48.3	62.3	56.0		89.0	74.0	70.1	68.6	80.0	67.9
8				80.0			85.0	65.0	65.0	78.7		77.0
9							83.3	80.0	95.0	95.0		89.4
10		55.0	45.0	50.0	62.5		73.5		80.0	64.0		67.7
11	73.2	65.7	45.9	61.8	56.4	71.0	79.3	67.8	66.4	68.4	67.5	67.1
12		67.6		63.1	56.0		73.1	70.0	67.5	75.8		68.9
13	80.0	66.6					77.2			57.5		73.2
14			45.0							85.0		71.6
22							65.0					65.0
25		70.4			79.4	80.0	83.0			82.5		77.5
27			45.0									45.0
28								50.0	76.6	90.0		75.4
29	65.0									65.0		65.0
30										65.0		65.0
31			45.0									45.0
32		80.0					84.1					83.5
35					50.0					95.0		72.5
36			50.0									50.0
38		65.0							65.0			65.0
39									50.0			50.0
40							80.0					80.0
43	95.0											95.0
44								65.0				65.0
45											65.0	65.0
AVERAGE	73.9	65.2	52.6	63.1	63.9	71.0	78.6	68.2	68.9	70.6	68.8	68.5

-THIS REPORT IS BASED ON A 10 YEARS TIME SPAN.
-BLANKS INDICATE NO MATCHING RECORDS.
-KEY TO MAJORS; 1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, 4 BTECH. ELECTRONICS, 5 BTECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, 8 B.SC. COMPUTERS, 9 B.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.
-KEY TO NATIONALITY; 1 JORDAN, 3 SUDAN, 4 SYRIA, 6 IRAQ, 7 PALESTINE, 8 QATAR, 9 LEBANON, 10 LIBYA, 11 EGYPT, 12 YEMEN, 12 KUWAIT, 14 ETHIOPIA, 22 NAMIBIA, 25 KSA, 27 GABON, 28 ERITREA, 29 OMAN, 30 KENYA, 31 ITALY, 32 UAE, 35 PAKISTAN, 36 CYPRUS, 38 TERKISTAN, 39 INDONESIA, 40 MOROCCO, 43 TUNISIA, 44 USA, 45 CANADA.

- **Years in University, Majors, and GPA**

User (s):

Dean	✓
Deputy Dean	✓
Registrar	
Admission Officer	
Other	

9/5/11

MAJORS, YEARS IN UNIVERSITY, AND GRADUATION GRADES

YEARS \ MAJORS	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	AVERAGES
3.0					57.5		65.0		65.0			62.5
4.0	84.1	79.7	62.5	80.0	68.0	76.3	78.5		74.0	48.3	53.7	78.2
5.0	67.9	68.2	54.4	65.9	72.5	69.1	83.2	77.2	74.3	77.9	76.0	72.6
6.0	58.0	53.4	49.0	61.4	58.6	50.0	74.6	66.1	67.3	73.3	59.0	66.3
7.0	50.0	51.9	48.5	61.5	59.0	50.0	67.8	54.8	58.8	57.3	65.0	56.8
8.0	80.0	50.0	55.0	61.1	52.5		70.0	57.5	55.4	63.3	65.0	59.2
9.0				61.2	60.0		57.5	50.0	53.7	64.7		60.8
10.0			50.0	72.5	50.0			50.0	65.0	56.5		58.0
11.0				50.0					80.0	80.0		70.0
12.0					65.0				50.0	65.0		60.0
13.0										50.0		50.0
AVERAGE	73.9	65.2	52.6	63.1	63.9	71.0	78.6	68.2	68.9	70.6	68.8	68.5

-BLANKS INDICATE NO MATCHING RECORDS.
-KEY TO MAJOR: 1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, 4 BTECH. ELECTRONICS, 5 BTECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, 8 B.SC. COMPUTERS, 9 B.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.

- **Majors versus Gender**

User (s):

Dean	✓
Deputy Dean	✓
Registrar	
Admission Officer	
Other	

MAJORS VERSUS GENDER

MAJORS:

BBA ENGLISH SECTION

TOTAL MAJOR 102

FEMALE	51.00	0.50	%
MALE	51.00	0.50	%

TOTAL GENDER 102

BBA ARABIC SECTION

TOTAL MAJOR 298

FEMALE	44.00	0.15	%
MALE	254.00	0.85	%

TOTAL GENDER 298

BACHELOR OF MARITIME TRANSPOR

TOTAL MAJOR 79

MALE	79.00	1.00	%

TOTAL GENDER 79

BTECH. ELECTRONICS

TOTAL MAJOR 99

MALE	99.00	1.00	%

TOTAL GENDER 99

BTECH. MARINE ENG.

TOTAL MAJOR 54

MALE	54.00	1.00	%

TOTAL GENDER 54

BACHELOR OF HOTELS AND TOURIS

MAJORS VERSUS GENDER

MAJORS:

TOTAL MAJOR 77

FEMALE	36.00	0.47	%
MALE	41.00	0.53	%

TOTAL GENDER 77

BACHELOR OF MARITIME

TOTAL MAJOR 118

MALE	118.00	1.00	%

TOTAL GENDER 118

B.SC. COMPUTERS

TOTAL MAJOR 185

FEMALE	32.00	0.17	%
MALE	153.00	0.83	%

TOTAL GENDER 185

B.SC. ELECTRONICS

TOTAL MAJOR 516

FEMALE	33.00	0.06	%
MALE	483.00	0.94	%

TOTAL GENDER 516

B.SC. MARINE ENG.

TOTAL MAJOR 433

MALE	433.00	1.00	%

TOTAL GENDER 433

MAJORS VERSUS GENDER

MAJORS:

B.SC. MECHANICAL ENG.

TOTAL MAJOR 39

FEMALE	1.00	0.03	%
MALE	38.00	0.97	%

TOTAL GENDER 39

GRAND TOTAL 2,000

-THIS REPORT IS BASED ON STUDENTS JOINED THE UNIVERSITY FROM 84.00 TO 93.00
; TIME SPAN IS 10.00 YEARS.

- **Age, GPA, and Majors**

User (s):

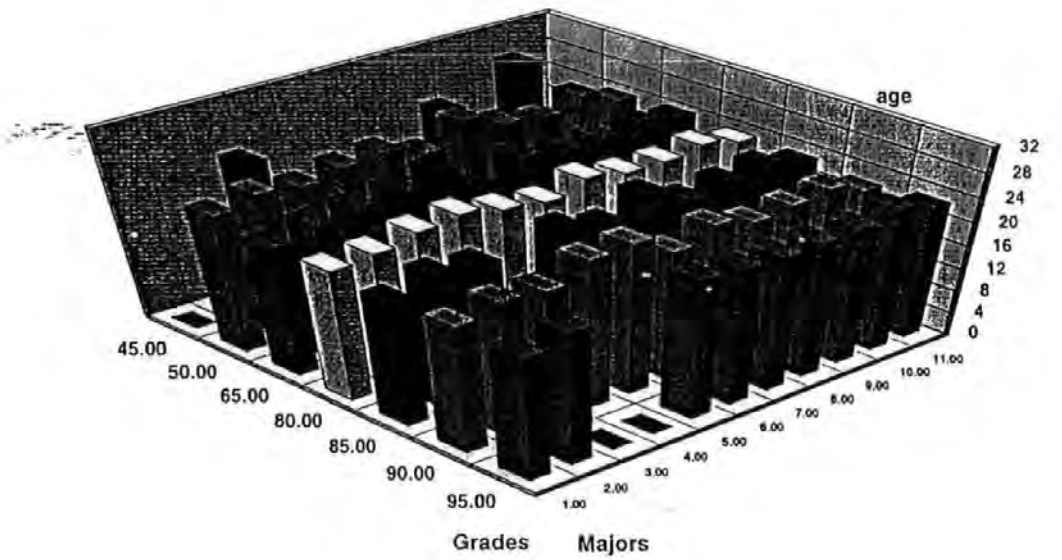
Dean	
Deputy Dean	
Registrar	
Admission Officer	✓
Other	

9/5/01

Majors, Graduation Grade, and Average Ages

%	Major											
	1	2	3	4	5	6	7	8	9	10	11	Averages
45.0			24.5					24.0		28.2		24.8
50.0	23.7	25.1	24.3	25.4	26.0	24.0	27.4	25.1	24.5	26.2	23.7	25.1
65.0	22.4	23.9	24.4	25.0	26.3	21.6	24.9	23.8	24.1	25.1	24.0	24.2
80.0	22.6	23.1	24.5	25.2	23.8	23.0	24.9	23.8	23.7	24.5	23.2	23.9
85.0	21.6	23.7	23.0		24.5	23.0	25.7	23.0	23.8	24.5	23.0	23.7
90.0	22.2	23.6	23.0	25.6	26.0	22.5	24.6	23.0	23.6	24.2	22.0	23.9
95.0	21.7	23.1			24.2	23.7	23.3	23.3	23.6	24.7	22.8	23.6
Averages	22.5	24.1	24.4	25.2	25.5	22.8	24.7	24.2	24.0	25.1	23.5	24.3

-blanks indicate no matching records.
-1 BBA English section, 2 BBA Arabic section, 3 Bachelor of Maritime Transport, 4 BTech. Electronics, 5 BTech. Marine Eng., 6 Bachelor of Hotels and Tourism, 7 Bachelor of Maritime, 8 B.Sc. Computers, 9 B.Sc. Electronics, 10 B.Sc. Marine Eng., 11 B.Sc. Mechanical Eng.



- **High Schools, Majors, and GPA**

User (s):

Dean	✓
Deputy Dean	
Registrar	✓
Admission Officer	
Other	✓

HIGH SCHOOLS, MAJORS, AND AVERAGE GRADUATION GRADES

H SCHOOLS	MAJOR											
	1	2	3	4	5	6	7	8	9	10	11	VERAGI
1	76.1	70.0	53.8	61.3	65.4	74.3	80.9	64.3	67.3	72.4	60.0	68.9
2	73.8	70.3	52.1	66.6	62.3	86.0	76.7	73.0	70.4	69.5	69.5	69.3
3	76.2	61.4		80.0	50.0	66.2		80.0		65.0		64.3
4		53.7										53.7
5	65.0	57.7										59.3
111				65.0					50.0			57.5
121		50.0										50.0
141		95.0		55.0	65.0			75.0	71.6			72.5
151		65.0										65.0
161								50.0	80.0			65.0
171								68.7	65.0			68.0
AVERAGE	73.9	65.2	52.6	63.1	63.9	71.0	78.6	68.2	68.9	70.6	68.8	68.5

-BLANKS INDICATE NO MATCHING RECORDS.
-KEY TO MAJORS; 1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, 4 BTECH. ELECTRONICS, 5 BTECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, 8 B.SC. COMPUTERS, 9 B.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.
-KEY TO HIGH SCHOOLS; 1 THANWYA AMMA-MATH, 2 THANWYA AMMA-SCIENCE, 3 THANWYA AZHAR, 4 PREPARATORY D 5 THANWYA AMMA-ARTS, 111 THANWYA AMMA-NEW, 121 THANWYA AMMA NEW- SCIENCE, 141 THANWYA AMMA OLD- SCIE 151 THANWYA AMMA OLD- ARTS, 161 THANWYA AMMA OLD- SCIENCE, 171 THANWYA AMMA OLD- MATH.

- **Majors value-added**

User (s):

Dean	
Deputy Dean	✓
Registrar	✓
Admission Officer	
Other	

THE MAJORS· VALUE-ADD

1 BBA ENGLISH SECTION	15.34
2 BBA ARABIC SECTION	8.97
3 BACHELOR OF MARITIME TRANSPORT	-8.51
4 BTECH. ELECTRONICS	-5.65
5 BTECH. MARINE ENG.	-1.52
6 BACHELOR OF HOTELS AND TOURISM	11.3
7 BACHELOR OF MARITIME	16.41
8 B.SC. COMPUTERS	-3.41
9 B.SC. ELECTRONICS	-2.78
10 B.SC. MARINE ENG.	3.67
11 B.SC. MECHANICAL ENG.	4.46

-THE HIGHEST VALUE-ADDED IS FOR THE 1 BBA ENGLISH SECTION 15.34;
-THE LOWEST VALUE-ADDED IS FOR THE 3 BACHELOR OF MARITIME TRANSPORT -8.51.

- **University value-added**

User (s):

Dean	✓
Deputy Dean	
Registrar	
Admission Officer	
Other	✓

THE UNIVERSITY VALUE-ADDED

UNIVERSITY VALUE-ADDED	2.71

THE AVERAGE HIGH SCHOOL PERCENTAGES	66.81
THE AVERAGE GRADUATION GRADES	68.56

- THE TIME SPAN ON THIS REPORT IS 10.00 YEARS
FOR STUDENTS WHO STARTED TO LEAVE THE UNIVERSITY IN 90.00 UP TO 99.00

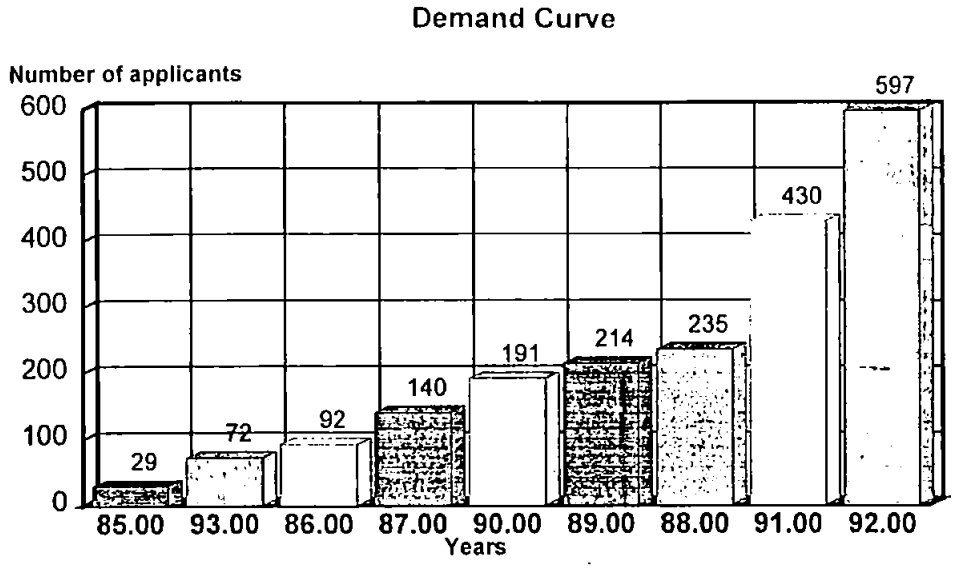
THE UNIVERSITY VALUE-ADDED VALUE IS CALCULATED AS FOLLOWS:
1-FIND THE STUDENTS· HIGH SCHOOL PERCENTAGES.
2-FIND THE STUDENTS· GRADUATION GRADES.
3-FIND THE DIFFERENCE BETWEEN THE TWO COLUMNS (I.E. GRADUATION GRADE - HIGH SCHOOL PERCENTAGE)
4-THE AVERAGE OF THE (GRADUATION GRADE - HIGH SCHOOL PERCENTAGE) COLUMN IS CONSIDERED THE UNIVERSITY VALUE-ADDED.

- **Demand Curve**

User (s):

Dean	✓
Deputy Dean	
Registrar	
Admission Officer	✓
Other	✓

THE UNIVERSITY DEMAND CURVE FROM 85 TO 93



YEAR JOINING THE UNIVERSITY

YEAR 85.00

TOTAL 29

YEAR 93.00

TOTAL 72

YEAR 86.00

TOTAL 92

YEAR 87.00

TOTAL 140

YEAR 90.00

TOTAL 191

YEAR 89.00

TOTAL 214

YEAR 88.00

TOTAL 235

YEAR 91.00

	YEAR JOINING THE UNIVERSITY
TOTAL	430
YEAR 92.00	
TOTAL	597
GRAND TOTAL	2,000

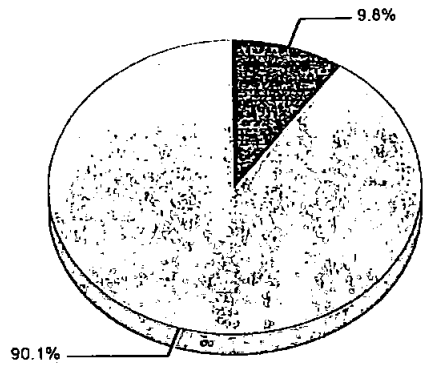
-THE TIME SPAN FOR THIS REPORT IS 9.00 YEARS
-THIS REPORTS DOES NOT REFLECT REALITY, AS THE DATA IS NOT COMPLETE; IT IS JUST 2000 RECORDS DRAWN AT RANDOM. HOWEVER, WHEN THIS REPORT RUNS ON A COMPLETE DATA SET, IT IS GOING TO MEET THE MANAGERS' NEEDS AND REFLECT THE REAL NUMVERS.

- **Gender Distribution**

User(s):

Dean	
Deputy Dean	
Registrar	
Admission Officer	✓
Other	✓

GENDER DISTRIBUTION



0.00

1.00

0 FOR FEMALE; 1 FOR MALE.

GENDER

FEMALES	0.00	
TOTAL		197
MALES	1.00	
TOTAL		1,803
GRAND TOTAL		2,000.00

- THE TIME SPAN ON THIS REPORT IS 10.00 YEARS
FOR STUDENTS WHO STARTED TO LEAVE THE UNIVERSITY IN 90.00 UP TO 99.00

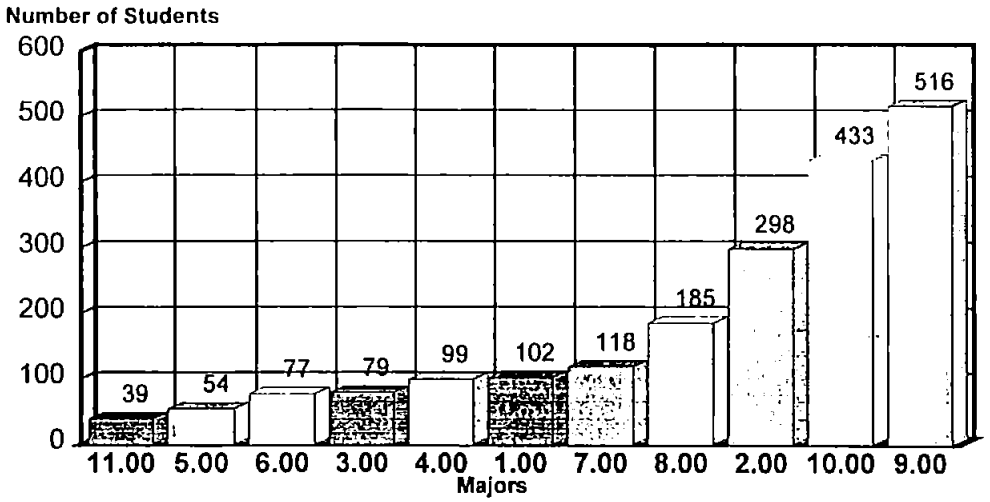
- **Major Distribution**

User (s):

Dean	
Deputy Dean	✓
Registrar	✓
Admission Officer	✓
Other	

MAJOR DISTRIBUTION

9.5.01



	MAJOR	
B.SC. MECHANICAL ENG.	11.00	
	TOTAL	39
BTECH. MARINE ENG.	5.00	
	TOTAL	54
BACHELOR OF HOTELS AND TO	6.00	
	TOTAL	77
BACHELOR OF MARITIME TRAN	3.00	
	TOTAL	79
BTECH. ELECTRONICS	4.00	
	TOTAL	99
BBA ENGLISH SECTION	1.00	
	TOTAL	102
BACHELOR OF MARITIME	7.00	
	TOTAL	118

MAJOR

B.SC. COMPUTERS	8.00	
	TOTAL	185
BBA ARABIC SECTION	2.00	
	TOTAL	298
B.SC. MARINE ENG.	10.00	
	TOTAL	433
B.SC. ELECTRONICS	9.00	
	TOTAL	516
	GRAND TOTAL	2,000.00

- THE TIME SPAN ON THIS REPORT IS 10.00 YEARS
FOR STUDENTS WHO STARTED TO LEAVE THE UNIVERSITY IN 90.00 UP TO 99.00
1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, 4 BTECH. ELECTR
5 BTECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, 8 B.SC. COMPU
9 B.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.

- **Applicants' High School Scores, Majors, and Average Graduation Scores**

User (s):

Dean	✓
Deputy Dean	
Registrar	
Admission Officer	✓
Other	

APPLICANTS' HIGH SCHOOL SCORES, MAJORS, AND AVERAGE GRADUATION GRADES

%	MAJORS											
	1	2	3	4	5	6	7	8	9	10	11	VERAGE
41							65.0					65.0
42		80.0										80.0
43		80.0	45.0							50.0		58.3
44							90.0		50.0			70.0
45		50.0					80.0			80.0		70.0
46		90.0								50.0		70.0
47		50.0					95.0					72.5
48		80.0			72.5		85.0		70.0	87.5	50.0	76.0
49							80.0		50.0	50.0		60.0
50	65.0	58.8	45.3	65.0	50.0	55.0	73.5	57.5	63.3	60.2	80.0	59.0
51	72.5	57.5	46.6	55.0	50.0	70.0	65.0	50.0	57.5	60.7		57.9
52		62.1	45.0	65.0		71.0	81.6	65.0	68.3	62.5	80.0	64.9
53	75.0	61.3	45.0		50.0	65.0	80.0		55.0	64.2		62.4
54	57.5	72.3	45.0			80.0	82.5	61.2	71.0	62.8		67.8
55	77.2	66.5	49.0	80.0	57.5	66.4	88.3	55.0	60.0	70.7	65.0	67.3
56	86.6	65.2	45.0	60.0	50.0	65.0	65.0	53.7	58.1	66.5	95.0	62.8
57	62.0	66.0	45.0	68.3	57.5	65.0	75.0	57.5	59.3	68.8	50.0	64.5
58	73.5	60.2		50.0		76.2	83.7	59.3	62.0	64.4	60.0	65.3
59	64.0	62.5	56.6	50.0	80.0	65.0	82.5	60.0	71.4	66.5	72.5	65.7
60	71.0	73.1	49.0	57.5	56.0	80.0	72.0	62.5	65.8	70.0	61.2	67.8
61	88.7	77.0		57.5		80.0	88.3	70.0	64.0	69.6	80.0	70.6
62	81.2	70.5	50.0	50.0	95.0	71.6	80.0	72.5	63.2	60.7	90.0	68.3
63	80.0	65.0	45.0	60.0	62.5	50.0	95.0	63.1	59.3	62.0	72.5	63.6
64	81.0	70.5	65.0	62.0	50.0		90.0	70.0	74.5	62.5	65.0	69.9
65	65.0	58.7	58.3	65.0		80.0	86.6	58.5	65.5	58.0		63.1
66	76.2	65.0	63.3		78.7	70.0	83.0	60.0	62.6	65.0		68.1
67	80.0	61.2	80.0	57.5		95.0	78.7	87.5	69.5	68.7		69.4
68	76.2	70.0	45.0	55.0	50.0		65.0	74.0	62.8	68.8	50.0	65.5
69	80.0	68.0	47.5		55.0			88.3	67.5	74.4	95.0	70.0
70		50.0	65.0	65.0	65.0	87.5	87.5	90.0	75.0	68.0	72.5	75.1
71	76.2	69.0	65.0	60.0	50.0			67.1	69.2	82.7	65.0	69.9
72		60.0		65.0	55.0		67.5	80.0	72.6	75.5	65.0	70.0
73	87.5	68.7	45.0	50.0	72.5	95.0	80.0	72.5	68.8	75.0	65.0	72.1
74		85.0	62.5	68.3		90.0	75.0	50.0	70.0	77.1		73.6
75	55.0	65.0	45.0	55.0	95.0		87.5	64.5	74.2	71.0		68.8
76		80.0	65.0	68.7	85.0		50.0	74.0	77.2	75.0	80.0	74.3
77		90.0	65.0	65.0	95.0	80.0	75.0	70.0	68.0	80.0		73.6
78	88.3	55.0	50.0	75.0				50.0	71.3	78.3	95.0	72.6
79				74.0	80.0	85.0		75.0	74.3	78.7		75.7
80	80.0	75.0		72.5	60.0	80.0		85.0	65.9	76.5		72.5
81			65.0					67.5	68.1	80.0	50.0	70.5
82		65.0	85.0	62.5	80.0		65.0	80.0	71.2	85.0	85.0	72.5
83		72.5		50.0	50.0	65.0		63.3	71.9	75.0		69.6
84							65.0	80.0	68.3	72.8		71.4
85				65.0	50.0	50.0	95.0	75.0	77.6	82.5		75.4
86		80.0	75.0	65.0	80.0			72.5	72.3	95.0		74.4

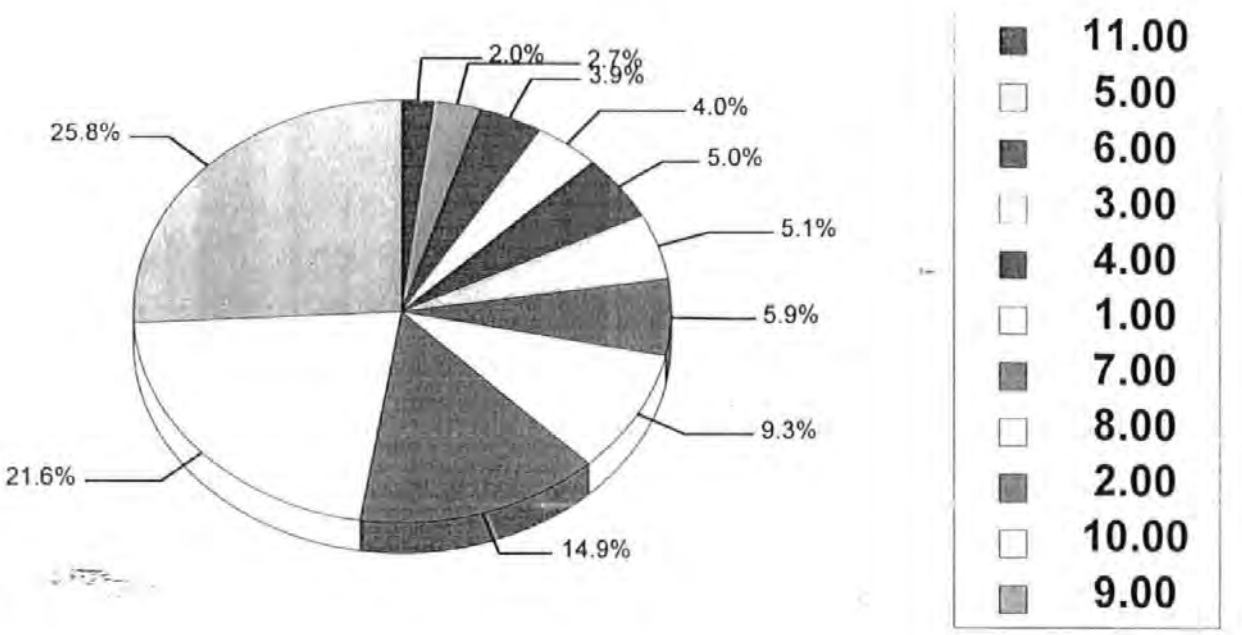
87	50.0	50.0	45.0	65.0		80.0	85.0	59.0	76.8	80.5		72.8
88		80.0		65.0	80.0		95.0	80.0	82.1	83.3		82.1
89	95.0			80.0	65.0			77.5	64.4			71.3
90	50.0			90.0				77.0	75.0	79.0		75.8
91					80.0			57.5	76.0	50.0		69.4
92	95.0			80.0			90.0	80.0	76.4	77.5		79.6
93								95.0	68.3	92.0		84.4
94									78.7	65.0	80.0	76.6
95								50.0		95.0		72.5
96									95.0			95.0
97	65.0											65.0
AVERAGES	73.9	65.2	52.6	63.1	63.9	71.0	78.6	68.2	68.9	70.6	68.8	68.5

-BLANKS INDICATE NO MATCHING RECORDS.
-KEY TO MAJORS: 1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, 4 BTECH. ELE
5 BTECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, 8 B.SC. COMPUTERS.
9 B.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.

4. The Visualization Reports

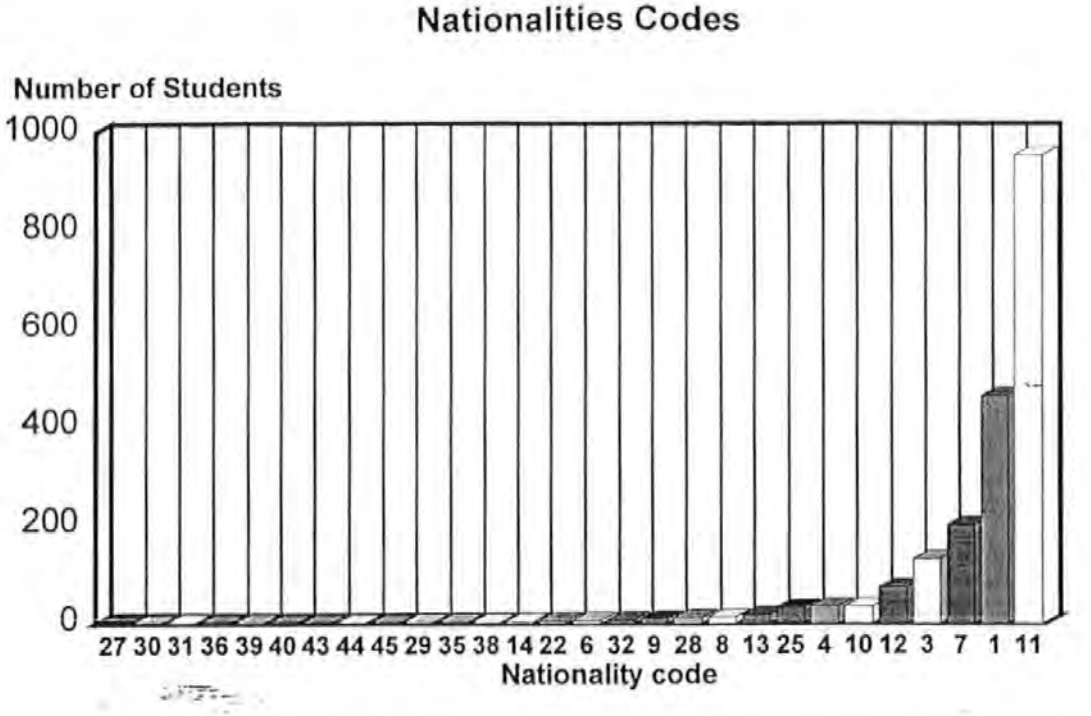
Visualization techniques have been utilized to extract another group of reports from the DW. These reports were generated to enable the Admission and Registration managers finding shallow and multi-dimensional knowledge and to represent general statistics. The number of such reports is dependent upon the managers' preferences and could vary from one University to another. No knowledge rules were derived based on these reports because different managers have different interpretations to the same report. Following are five example reports.

VISUALIZATION (1)

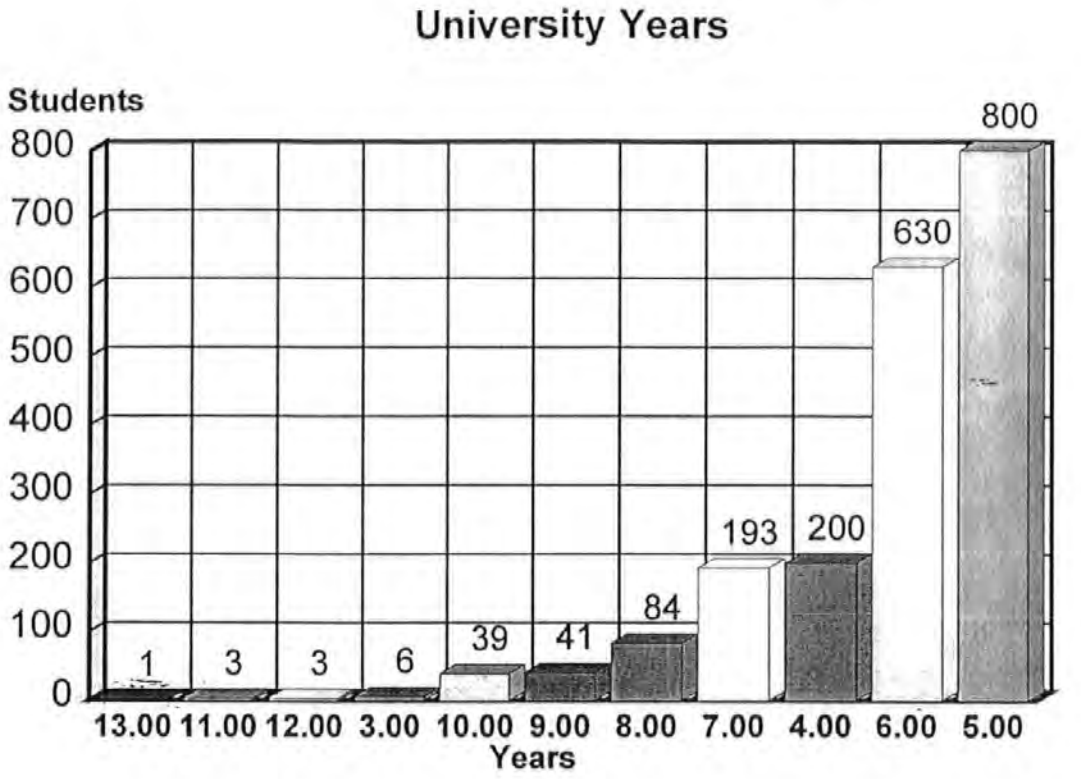


Y TO MAJORS: 1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, TECH. ELECTRONICS, 5 BTECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, SC. COMPUTERS, 9 B.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.

VISUALIZATION (2)

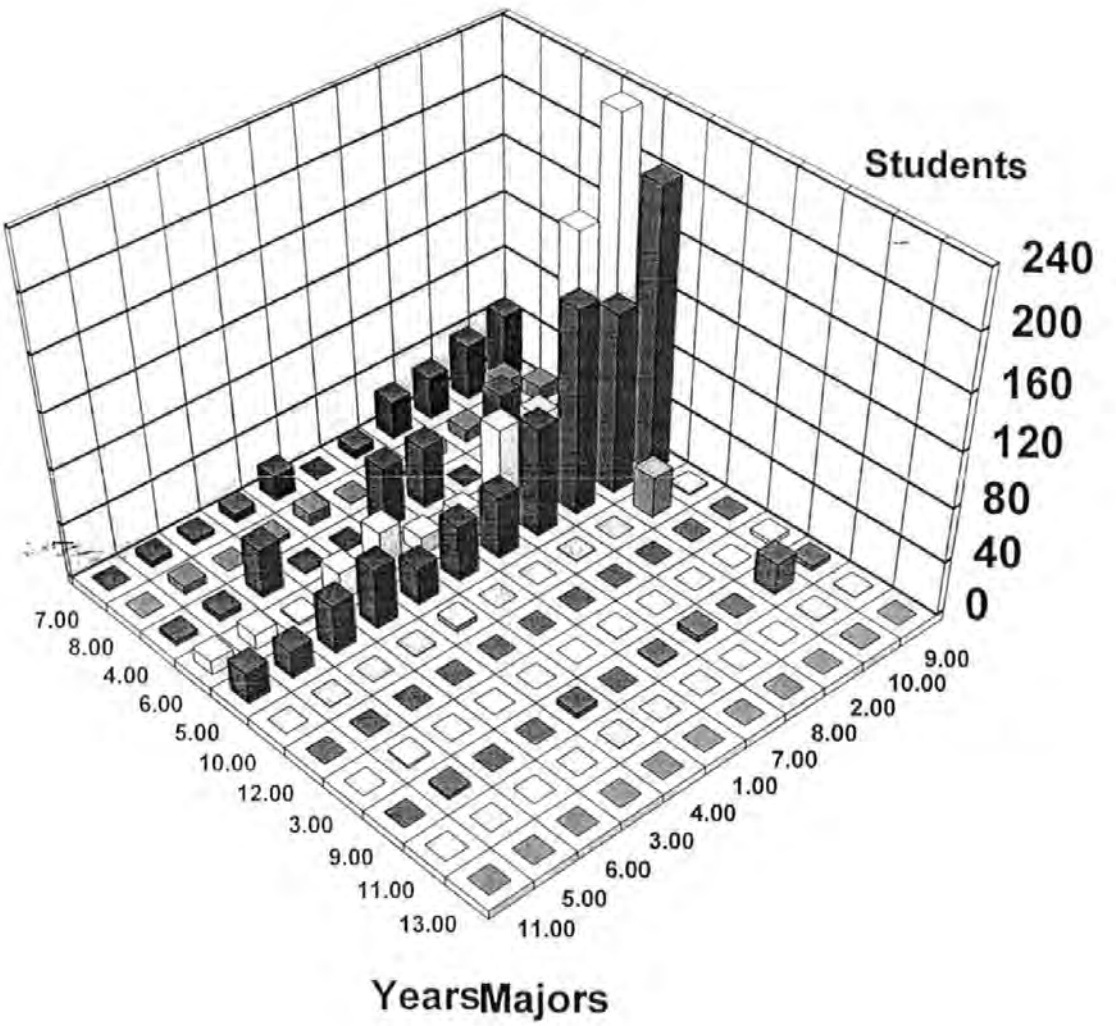


EY TO NATIONALITY; 1 JORDAN, 3 SUDAN, 4 SYRIA, 6 IRAQ, 7 PALESTINE, 8 QATAR, 9 LEBANON, 10 LIBYA, 11 EGYPT, YEMEN, 12 KUWAIT, 14 ETHIOPIA, 22 NAMIBIA, 25 KSA, 27 GABON, 28 ERITREA, 29 OMAN, 30 KENYA, 31 ITALY, UAE, 35 PAKISTAN, 36 CYPRUS, 38 TERKISTAN, 39 INDONESIA, 40 MOROCCO, 43 TUNISIA, 44 USA, 45 CANADA.

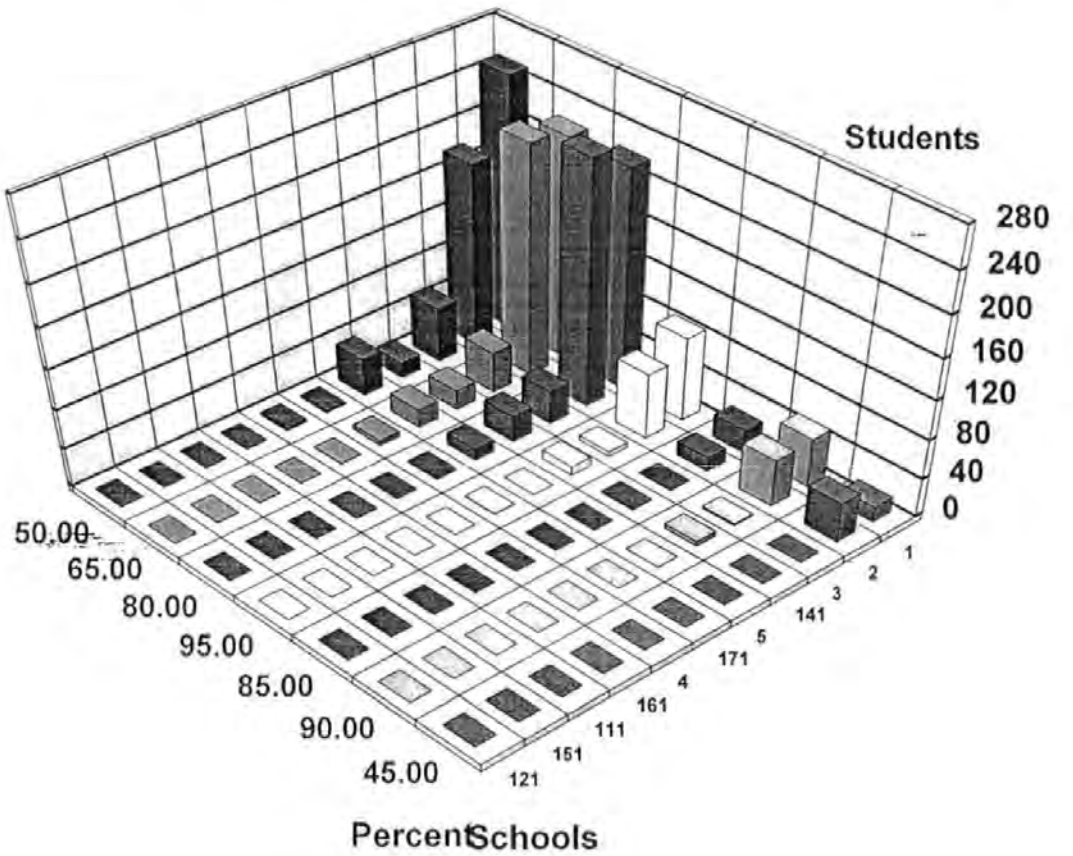

University Years
Students
800
700
600
500
400
300
200
100
0
1
3
3
6
39
41
84
193
200
630
800
13.00
11.00
12.00
3.00
10.00
9.00
8.00
7.00
4.00
6.00
5.00
Years

VISUALIZATION (4)

University Years and Majors



Y TO MAJORS: 1 BBA ENGLISH SECTION, 2 BBA ARABIC SECTION, 3 BACHELOR OF MARITIME TRANSPORT, 4 BTECH. ELE
TECH. MARINE ENG., 6 BACHELOR OF HOTELS AND TOURISM, 7 BACHELOR OF MARITIME, 8 B.SC. COMPUTERS,
.SC. ELECTRONICS, 10 B.SC. MARINE ENG., 11 B.SC. MECHANICAL ENG.



EY TO HIGH SCHOOLS; 1 THANWYA AMMA-MATH, 2 THANWYA AMMA-SCIENCE, 3 THANWYA AZHAR, REPARATORY DIPLOMA, 5 THANWYA AMMA-ARTS, 111 THANWYA AMMA-NEW, 121 THANWYA AMMA NEW- SCIENCE, THANWYA AMMA OLD- SCIENCE, 151 THANWYA AMMA OLD- ARTS, 161 THANWYA AMMA OLD- SCIENCE, THANWYA AMMA OLD- MATH.

Appendix (E)

Proximity Measures' Calculations

1-The first group calculations (similar records)

1-1Gower similarity coefficient: $S_{12} = .9975$

Student 1	Student 2	Range	S_{12}
1	1	1	1
11	11	11	1
92	92	3	1
92	92	3	1
65	66	37	0.973
8	8	8	1
98	98	3	1
1	1	5	1
74	74	19	1
11	11	10	1
1	1	1	1
			10.9725
		10.9725/11	**0.9975**

Table (E-1). Explanation of Gower.

1-2 Euclidean metric measure: $D_{12} = 1$

Student 1	Student 2	Student 1- Student 2	(Student 1 - Student 2)^2
1	1	0	0
11	11	0	0
92	92	0	0
92	92	0	0
65	66	-1	1
8	8	0	0
98	98	0	0
1	1	0	0
74	74	0	0

11	11	0	0
1	1	0	0
			1
		$\sqrt{1}$	**1**

Table (E-2). Explanation of Euclidean.

1-3 Modified Euclidean metric measure: $D_{12} = .08475$

Student 1	Student 2	σ	Z_1	Z_2	$Z_1 - Z_2$	$(Z_1 - Z_2)^2$
1	1	0.38	2.632	2.632	0	0
11	11	5.88	1.871	1.871	0	0
92	92	1.13	81.416	81.416	0	0
92	92	1.13	81.416	81.416	0	0
65	66	11.8	5.508	5.593	-0.085	0.0072
8	8	3.3	2.424	2.424	0	0
98	98	1.15	85.217	85.217	0	0
1	1	2.56	0.391	0.391	0	0
74	74	6.26	11.821	11.821	0	0
11	11	3.78	2.910	2.910	0	0
1	1	0.53	1.887	1.887	0	0
						0.0072
					$\sqrt{.0072}$	**.08475**

Table (E-3). Explanation of modified Euclidean.

1-4 City Block Metric measure: $D_{12} = 1$

Student 1	Student 2	$\lvert X_{ik} - X_{jk} \rvert$
1	1	0
11	11	0
92	92	0
92	92	0
65	66	1

8	8	0
98	98	0
1	1	0
74	74	0
11	11	0
1	1	0
		1

Table (E-4). Explanation of City Block Metric.

1-5 Canberra Metric measure: D_{12} = .007

Student 1	**Student 2**	$\lvert X_{ik} - X_{jk} \rvert$	$(X_{ik} + X_{jk})$	**D_{12}**
1	1	0	2	0
11	11	0	22	0
92	92	0	184	0
92	92	0	184	0
65	66	1	131	.007
8	8	0	16	0
98	98	0	196	0
1	1	0	2	0
74	74	0	148	0
11	11	0	22	0
1	1	0	2	0
				.007

Table (E-5). Explanation of Canberra Metric.

2-The second group calculations (dissimilar records)

2-1 Gower similarity coefficient: S_{46} = .21197

Student 4	**Student 6**	**Range**	**S_{46}**
1	2	1	0
0	11	11	0

89	92	3	0
89	92	3	0
76	64	37	0.676
10	2	8	0.000
95	96	3	0.667
2	6	5	0.200
71	75	19	0.789
1	11	10	0.000
1	0	1	0.000
			2.33169
			.21197

Table (E-6). Explanation of Gower-1.

2-2 Euclidean metric measure: $D_{46} = 21.9545$

Student 4	Student 6	Student 4- Student 6	(Student 4 - Student 6)^2
1	2	-1	1
0	11	-11	121
89	92	-3	9
89	92	-3	9
76	64	12	144
10	2	8	64
95	96	-1	1
2	6	-4	16
71	75	-4	16
1	11	-10	100
1	0	1	1
			482
		$\sqrt{482}$	**21.9545**

Table (E-7). Explanation of Euclidean-1.

2-3 Modified Euclidean metric measure: D_{46} = 6.75257

Student 4	Student 6	σ	Z_4	Z_6	$Z_4 - Z_6$	$(Z_4 - Z_6)^2$
1	2	0.38	2.632	5.263	-2.632	6.925
0	11	5.88	0.000	1.871	-1.871	3.500
89	92	1.13	78.761	81.416	-2.655	7.048
89	92	1.13	78.761	81.416	-2.655	7.048
76	64	11.8	6.441	5.424	1.017	1.034
10	2	3.3	3.030	0.606	2.424	5.877
95	96	1.15	82.609	83.478	-0.870	0.756
2	6	2.56	0.781	2.344	-1.563	2.441
71	75	6.26	11.342	11.981	-0.639	0.408
1	11	3.78	0.265	2.910	-2.646	6.999
1	0	0.53	1.887	0.000	1.887	3.560
						45.5972
					$\sqrt{45.5972}$	**6.75257**

Table (E-8). Explanation of modified Euclidean-1.

2-4 City Block Metric measure: D_{46} = 58

Student 4	Student 6	$\lvert X_{ik} - X_{jk} \rvert$
1	2	1
0	11	11
89	92	3
89	92	3
76	64	12
10	2	8
95	96	1
2	6	4
71	75	4
1	11	10

1	0	1
		58

Table (E-9). Explanation of City Block Metric-1.

2-5 Canberra Metric measure: D_{46} = 4.46

Student 4	**Student 6**	$\lvert X_{ik} - X_{jk} \rvert$	$(X_{ik} + X_{jk})$	**D_{46}**
1	2	1	3	.33
0	11	11	11	1
89	92	3	181	.016
89	92	3	181	.016
76	64	12	140	.086
10	2	8	12	.66
95	96	1	191	.005
2	6	4	8	.50
71	75	4	146	.027
1	11	10	12	.83
1	0	1	1	1
				4.47

Table (E-10). Explanation of Canberra Metric-1.

3-The third group calculations (fairly similar records)

3-1 Similar in seven, different in four variables

3-1-1 Gower similarity coefficient: S_{35} = .7175

Student 3	**Student 5**	**Range**	**S_{35}**
1	1	1	1
11	0	11	0
92	92	3	1
92	92	3	1

75	54	37	0.432432432
8	2	8	0.25
97	97	3	1
6	6	5	1
75	90	19	0.210526316
11	11	10	1
0	0	1	1
			7.8929
			0.7175

Table (E-11). Explanation of Gower-2.

3-1-2 Euclidean metric measure: $D_{35} = 28.6879$

Student 3	**Student 5**	**Student 3- Student 5**	**(Student 3 - Student 5)^2**
1	1	0	0
11	0	11	121
92	92	0	0
92	92	0	0
75	54	21	441
8	2	6	36
97	97	0	0
6	6	0	0
75	90	-15	225
11	11	0	0
0	0	0	0
			823
		$\sqrt{823}$	**28.68798**

Table (E-12). Explanation of Euclidean-2.

3-1-3 Modified Euclidean metric measure: $D_{35} = 3.9641$

Student 3	Student 5	σ	Z_3	Z_5	$Z_3 - Z_5$	$(Z_3 - Z_5)^2$
1	1	0.38	2.632	2.632	0.000	0.000
11	0	5.88	1.871	0.000	1.871	3.500
92	92	1.13	81.416	81.416	0.000	0.000
92	92	1.13	81.416	81.416	0.000	0.000
75	54	11.8	6.356	4.576	1.780	3.167
8	2	3.3	2.424	0.606	1.818	3.306
97	97	1.15	84.348	84.348	0.000	0.000
6	6	2.56	2.344	2.344	0.000	0.000
75	90	6.26	11.981	14.377	-2.396	5.742
11	11	3.78	2.910	2.910	0.000	0.000
0	0	0.53	0.000	0.000	0.000	0.000
						15.7143
					$\sqrt{15.7143}$	**3.9641**

Table (E-13). Explanation of modified Euclidean-2.

3-1-4 City Block Metric measure: $D_{35} = 53$

Student 3	Student 5	$\lvert X_{ik} - X_{jk} \rvert$
1	1	0
11	0	11
92	92	0
92	92	0
75	54	21
8	2	6
97	97	0
6	6	0
75	90	15
11	11	0

0	0	0
		53

Table (E-14). Explanation of City Block Metric-2.

3-1-5Canberra Metric measure: D_{35} = 1.852

Student 3	**Student 5**	$\lvert X_{ik} - X_{jk} \rvert$	$(X_{ik} + X_{jk})$	**D_{35}**
1	1	0	2	0
11	0	11	11	1
92	92	0	184	0
92	92	0	184	0
75	54	21	129	.162
8	2	6	10	.6
97	97	0	194	0
6	6	0	12	0
75	90	15	165	.090
11	11	0	22	0
0	0	0	0	0
				1.852

Table (E-15). Explanation of Canberra Metric-2.

3-2 Similar in four, different in seven variables (opposite of the last case)

3-2-1 Gower similarity coefficient: S_{57} = .5106

Student 5	**Student 7**	**Range**	**S_{57}**
1	1	1	1
0	0	11	1
92	91	3	0.6667
92	92	3	1
54	91	37	0.000

2	9	8	0.125
97	98	3	0.667
6	1	5	0.000
90	74	19	0.158
11	11	10	1.000
0	1	1	0.000
			5.6162
			.5106

Table (E-16). Explanation of Gower-3.

3-2-2 Euclidean metric measure: $D_{57} = 41.2553$

Student 5	**Student 7**	**Student 5- Student 7**	**(Student 5 - Student 7)^2**
1	1	0	0
0	0	0	0
92	91	1	1
92	92	0	0
54	91	-37	1369
2	9	-7	49
97	98	-1	1
6	1	5	25
90	74	16	256
11	11	0	0
0	1	-1	1
			1702
		$\sqrt{1702}$	**41.2553**

Table (E-17). Explanation of Euclidean-3.

3-2-3 Modified Euclidean metric measure: $D_{57} = 5.4569$

Student 5	Student 7	σ	Z_5	Z_7	$Z_5 - Z_7$	$(Z_5 - Z_7)^2$
1	1	0.38	2.632	2.632	0	0
0	0	5.88	0.000	0.000	0	0
92	91	1.13	81.416	80.531	0.885	0.7831
92	92	1.13	81.416	81.416	0	0
54	91	11.8	4.576	7.712	-3.136	9.8319
2	9	3.3	0.606	2.727	-2.121	4.4995
97	98	1.15	84.348	85.217	-0.87	0.7561
6	1	2.56	2.344	0.391	1.953	3.8147
90	74	6.26	14.377	11.821	2.556	6.5327
11	11	3.78	2.910	2.910	0	0
0	1	0.53	0.000	1.887	-1.887	3.56
						29.778
					$\sqrt{29.778}$	**5.4569**

Table (E-18). Explanation of modified Euclidean-3.

3-2-4 City Block Metric measure: $D_{57} = 68$

Student 5	Student 7	$\lvert X_{ik} - X_{jk} \rvert$
1	1	0
0	0	0
92	91	1
92	92	0
54	91	37
2	9	7
97	98	1
6	1	5
90	74	16
11	11	0

0	1	1
		68

Table (E-19). Explanation of City Block Metric-3.

3-2-5 Canberra Metric measure: $D_{57} = 2.682$

Student 5	**Student 7**	$\lvert X_{ik} - X_{jk} \rvert$	$(X_{ik} + X_{jk})$	$\mathbf{D_{57}}$
1	1	0	2	0
0	0	0	0	0
92	91	1	183	.005
92	92	0	184	0
54	91	37	145	.225
2	9	7	11	.636
97	98	1	195	.005
6	1	5	7	.714
90	74	16	164	.097
11	11	0	22	0
0	1	1	1	1
				2.682

Table (E-20). Explanation of Canberra Metric-3.

Appendix (F)

The ARDSS technical documents

1- CLUSTAN outputs

1-1 Decision number F, using ISS (Ward)

1-1-1 Clusters' members

Cluster	Members
Cluster 1	BBA-EN [38]
Cluster 2	BBA-EN [81]
Cluster 3	BTECH-MREN [136]
Cluster 4	BBA-EN [50]
Cluster 5	BBA-EN [146]
Cluster 6	BBA-EN [60]
Cluster 7	BBA-AR [42]
Cluster 8	BBA-EN [46]
Cluster 9	BBA-AR [32]
Cluster 10	BMTRM [66]
Cluster 11	BTECH-ELEC [206]
Cluster 12	BTECH-ELEC [89]
Cluster 13	BBA-EN [78]
Cluster 14	BTECH-MREN [78]
Cluster 15	BBA-EN [37]
Cluster 16	BBA-AR [128]
Cluster 17	BBA-EN [92]
Cluster 18	BSC-COMP [58]
Cluster 19	BBA-EN [107]
Cluster 20	BTECH-MREN [14]
Cluster 21	BBA-EN [133]
Cluster 22	BBA-AR [9]
Cluster 23	BSC-COMP [27]
Cluster 24	BTECH-ELEC [37]
Cluster 25	BSC-COMP [10]

Table (F-1). Clusters' members for decision (F).

1-1-2Clusters' Dendrogram

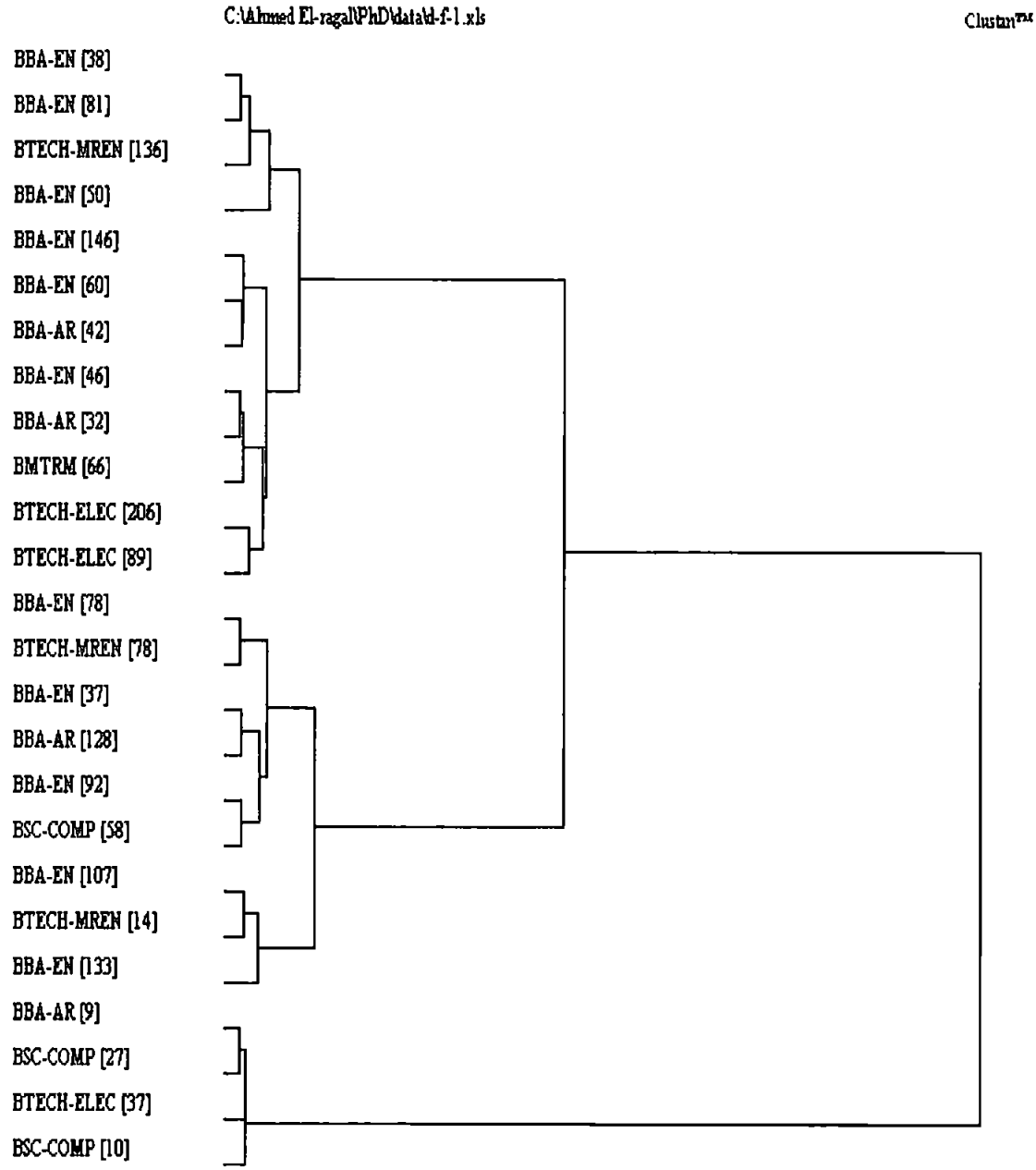


Figure (F-1). Dendrogram for decision (F).

1-1-3 Clusters' profiles

Clusters	High School	High School origin	High School percent	Major	GPA	Nationality	Gender	Age on graduation	Members
Cluster 1	3	11	66	2	3	10	1	23	38
Cluster 2	1	11	65	9	2	11	1	25	81
Cluster 3	2	11	65	6	3	29	1	27	136
Cluster 4	2	11	56	2	2	11	1	24	50
Cluster 5	3	11	52	3	1	11	1	25	146
Cluster 6	2	11	52	4	9	11	1	25	60
Cluster 7	2	11	55	3	2	3	1	25	42
Cluster 8	1	11	50	9	2	4	1	26	46
Cluster 9	2	11	59	10	2	2	1	26	32
Cluster 10	2	11	58	9	3	11	1	25	66
Cluster 11	2	11	51	8	2	11	1	25	206
Cluster 12	1	11	78	7	3	9	1	24	89
Cluster 13	1	11	73	9	3	11	1	24	78
Cluster 14	1	11	66	3	2	2	1	25	78
Cluster 15	2	12	68	9	2	3	1	25	37
Cluster 16	1	10	74	8	3	1	1	25	128
Cluster 17	2	11	80	9	3	1	1	24	92
Cluster 18	1	11	87	8	3	10	1	24	58
Cluster 19	1	11	85	7	4	25	1	25	107
Cluster 20	1	11	86	8	3	1	1	24	14
Cluster 21	133	28	76	6	2	11	1	25	133
Cluster 22	141	9	80	8	3	9	1	25	9
Cluster 23	141	10	63	8	4	9	1	23	27
Cluster 24	171	11	69	8	2	10	1	25	37
Cluster 25									10

Table (F-2). Clusters' profiles for decision (F).

1-2 Decision number L, using ISS (Ward)

1-2-1 Clusters' members

Cluster	Members
Cluster 1	BBA-EN [44]
Cluster 2	BBA-EN [141]
Cluster 3	BBA-EN [128]
Cluster 4	BMTRM [65]
Cluster 5	BBA-EN [116]
Cluster 6	BSC-COMP [84]
Cluster 7	BBA-EN [79]
Cluster 8	BTECH-MREN [148]
Cluster 9	BHOTORM [103]
Cluster 10	BBA-EN [147]
Cluster 11	BBA-EN [115]
Cluster 12	BBA-AR [92]
Cluster 13	BMTRSP [45]
Cluster 14	BTECH-ELEC [60]
Cluster 15	BBA-EN [43]
Cluster 16	BBA-AR [21]
Cluster 17	BBA-EN [75]
Cluster 18	BMTRM [42]
Cluster 19	BSC-COMP [74]
Cluster 20	BTECH-ELEC [43]
Cluster 21	BHOTORM [52]
Cluster 22	BBA-AR [9]
Cluster 23	BSC-COMP [30]
Cluster 24	BTECH-ELEC [34]
Cluster 25	BSC-COMP [10]

Table (F-3). Clusters' members for decision (L).

1-2-2 Clusters' Dendrogram

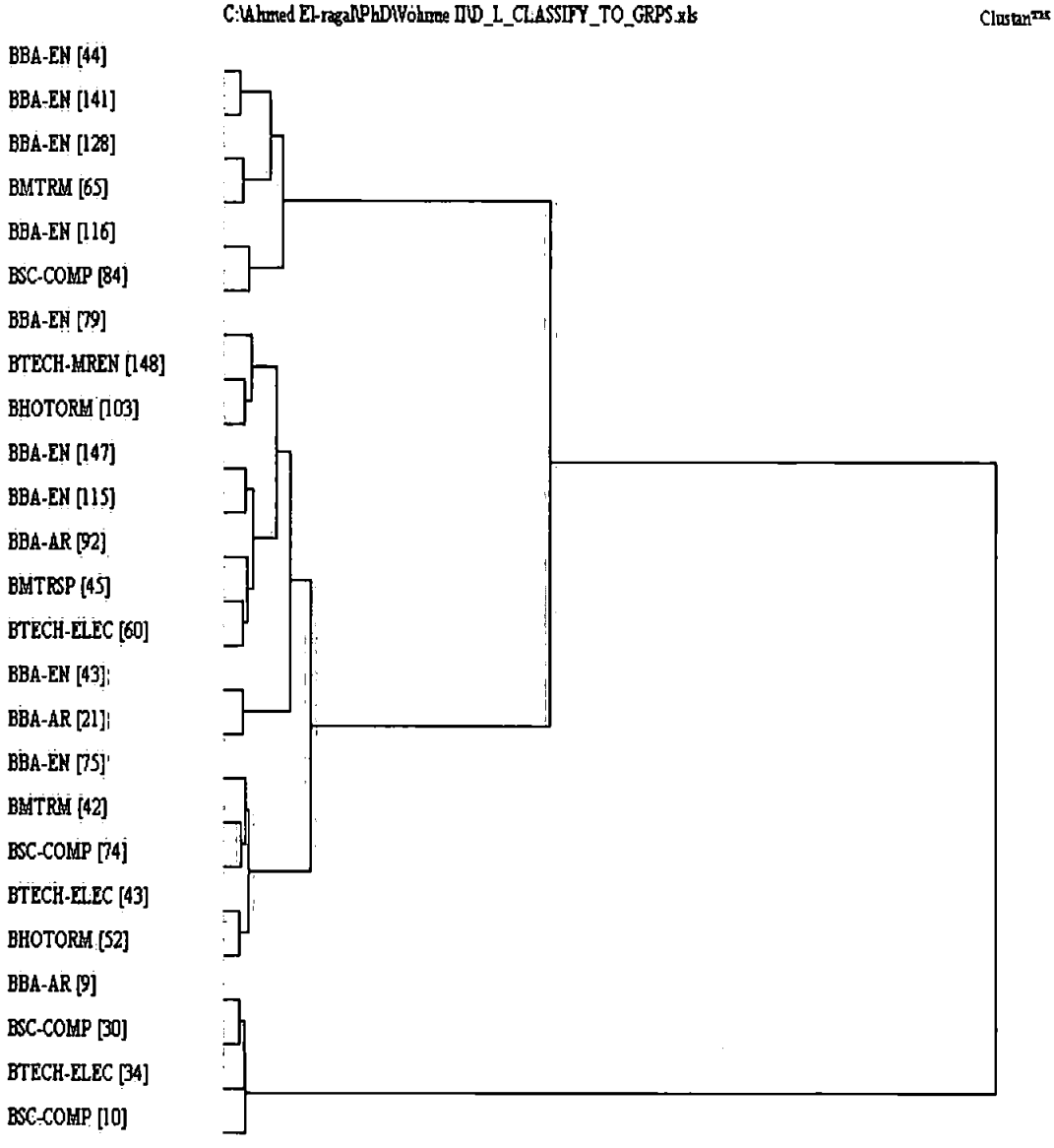


Figure (F-2). Dendrogram for decision (L).

1-2-3 Clusters' profiles

Clusters	HiSchlCod	HiSchlNation	HiSchlYr	JoinUnivYr	HiSchlPr%	GrdtMjr	GrdtnDate	GrdtnGrd	DoB	NtnltyCod	Gender	Members
Cluster 1	2	11	90	91	73	3	96	3	72	10	1	44
Cluster 2	1	11	91	91	75	9	97	3	73	11	1	141
Cluster 3	1	11	88	88	80	8	94	3	70	1	1	128
Cluster 4	2	11	88	88	73	10	95	3	70	1	1	65
Cluster 5	1	11	90	90	88	8	96	3	72	9	1	116
Cluster 6	2	11	88	89	88	9	94	4	70	1	1	84
Cluster 7	2	11	91	91	64	2	96	3	75	11	2	79
Cluster 8	1	11	91	91	63	8	97	2	73	11	1	148
Cluster 9	2	11	92	92	57	8	97	2	74	11	1	103
Cluster 10	2	11	91	91	56	2	96	3	73	11	1	147
Cluster 11	3	11	91	91	51	4	97	2	73	11	1	115
Cluster 12	2	10	90	90	55	6	97	2	72	4	1	92
Cluster 13	2	11	87	88	55	3	94	8	69	10	1	45
Cluster 14	1	11	88	88	53	9	95	2	70	11	1	60
Cluster 15	1	11	89	90	74	6	96	3	70	29	1	43
Cluster 16	1	11	88	89	57	7	95	3	66	28	1	21
Cluster 17	2	11	90	90	68	5	96	2	72	3	1	75
Cluster 18	2	11	87	87	58	10	93	2	68	1	1	42
Cluster 19	2	11	87	87	66	10	94	2	68	1	1	74
Cluster 20	2	11	87	87	65	9	94	3	68	11	1	95
Cluster 21	121	28	90	90	76	6	97	2	72	12	1	9
Cluster 22	141	9	91	91	79	9	97	3	74	9	2	30
Cluster 23	141	10	91	91	62	8	97	2	73	10	1	34
Cluster 24	169	11	91	91	69	8	97	2	73	10	1	10
Cluster 25												

Table (F-4). Clusters' profiles for decisions (L).

2- The ARDSS[1] DB CREATE statements

```
  USE master
GO
  CREATE DATABASE IEFDB
    ON DEFAULT =       20
GO
  USE IEFDB
GO
  CREATE TABLE  APPLICANT
    (APP_FULL_NAME                    char(35)            NOT NULL,
     APP_DOB                          datetime            NOT NULL,
     APP_GENDER                       char(1)
       DEFAULT 'm'
       NULL ,
     APP_ADDRESS                      char(35)            NULL,
     APP_PREVIOUSLY_ABANDONED         char(1)
       DEFAULT 'n'
       NULL ,
     APP_TRANSFERRED                  char(1)
       DEFAULT 'n'
       NULL ,
     APP_CODE                         int                 NOT NULL,
     APP_CERTIFICATE_YEAR             int                 NOT NULL,
     APP_CERTIFICATION_PERCENT        float               NOT NULL,
     APP_TELEPHONE                    char(15)            NULL,
     APP_INTERVIEWED                  char(1)
       DEFAULT 'p'
       NULL ,
     FK_UNIVERSITYUNI_ID              int                 NULL,
     FK_BATCHBAT_NUMBER               int                 NULL,
     FK_MAJORMAJ_SNO                  int                 NULL,
     FK_CERTIFICATECER_NUMBER         int                 NULL,
     FK_NATIONALITYNAT_IDENTIFICATI   int                 NULL,
    CONSTRAINT appcode
    PRIMARY KEY NONCLUSTERED
      (APP_CODE                         ))
GO
  CREATE TABLE  ASSISTANTSHIP
    (ASS_TITLE                        char(30)            NOT NULL,
     ASS_NUMBER                       int                 NOT NULL,
     ASS_REQUIREMENTS                 char(30)            NULL,
     ASS_DISCOUNT_RATE                float
       DEFAULT 25
       NOT NULL ,
     ASS_CATEGORY                     char(25)
       DEFAULT 'academic'
       NULL ,
    CONSTRAINT assnumber
    PRIMARY KEY NONCLUSTERED
      (ASS_NUMBER                       ))
GO
  CREATE TABLE  AUTHORITY
    (AUT_NAME                         char(25)            NOT NULL,
     AUT_CODE                         int                 NOT NULL,
     AUT_ADDRESS                      char(35)            NULL,
     AUT_TELEPHONE                    char(12)            NULL,
     AUT_CONTACT_PERSON               char(25)            NULL,
    CONSTRAINT autcode
```

[1] The **ARDSS** source code exceeds **68000** lines of **C Language**, so it will not be listed here for space reasons.

```
    PRIMARY KEY NONCLUSTERED
      (AUT_CODE                          ))
GO
  CREATE TABLE  BATCH
    (BAT_TITLE                          char(10)
       DEFAULT 'September'
       NULL ,
     BAT_NUMBER                         int                 NOT NULL,
     BAT_CEILING                        int                 NOT NULL,
     BAT_YEAR                           int                 NOT NULL,
     BAT_OPEN_DATE                      datetime            NULL,
     BAT_CLOSING_DATE                   datetime            NOT NULL,
     BAT_MARKET_SHARE                   float               NULL,
     BAT_GOVERNMENT_STAT                int                 NULL,
     FK_SEMESTERSEM_CODE                int                 NULL,
    CONSTRAINT batnumber
    PRIMARY KEY NONCLUSTERED
      (BAT_NUMBER                        ))
GO
  CREATE TABLE  CERTIFICATE
    (CER_NUMBER                         int                 NOT NULL,
     CER_NAME                           char(35)            NOT NULL,
     CER_ORIGIN                         char(25)
       DEFAULT 'egypt'
       NULL ,
    CONSTRAINT cernumber
    PRIMARY KEY NONCLUSTERED
      (CER_NUMBER                        ))
GO
  CREATE TABLE  COLLEGE
    (COL_NAME                           char(35)            NOT NULL,
     COL_SERIAL                         int                 NOT NULL,
     COL_LOCATION                       char(25)            NULL,
    CONSTRAINT colserial
    PRIMARY KEY NONCLUSTERED
      (COL_SERIAL                        ))
GO
  CREATE TABLE  COURSE
    (COU_STAGE                          char(15)            NULL,
     COU_TITLE                          char(25)            NOT NULL,
     COU_CODE                           char(8)             NOT NULL,
     COU_CREDIT_HOURS                   float
       DEFAULT 3
       NOT NULL ,
     COU_PASS_MARK                      float
       DEFAULT 50
       NOT NULL ,
     COU_FULL_MARK                      float
       DEFAULT 100
       NOT NULL ,
     COU_AREA                           char(25)            NULL,
    CONSTRAINT coucode
    PRIMARY KEY NONCLUSTERED
      (COU_CODE                          ))
GO
  CREATE TABLE  COURSE_MAJOR
    (COU_MAJ_COUNTER                    int                 NOT NULL,
     FK_COURSECOU_CODE                  char(8)             NULL,
     FK_MAJORMAJ_SNO                    int                 NULL,
    CONSTRAINT coumajcounter
    PRIMARY KEY NONCLUSTERED
```

```
      (COU_MAJ_COUNTER                    ))
GO
  CREATE TABLE  DEPARTMENT
    (DEP_TITLE                      char(20)            NOT NULL,
     DEP_ID                         int                 NOT NULL,
     DEP_LOCATION                   char(25)            NULL,
     DEP_TYPE                       char(15)
       DEFAULT 'education'
       NULL ,
     FK_COLLEGECOL_SERIAL           int                 NULL,
    CONSTRAINT depid
    PRIMARY KEY NONCLUSTERED
      (DEP_ID                         ))
GO
  CREATE TABLE  EXAM
    (EXA_NUMBER                     int                 NOT NULL,
     EXA_TYPE                       char(15)            NOT NULL,
     FK_COURSECOU_CODE              char(8)             NULL,
    CONSTRAINT exanumber
    PRIMARY KEY NONCLUSTERED
      (EXA_NUMBER                     ))
GO
  CREATE TABLE  GPA
    (GPA_COUNTER                    int                 NOT NULL,
     FK_STUDENTSTU_REGISTRATION_NUM float               NULL,
    CONSTRAINT gpacounter
    PRIMARY KEY NONCLUSTERED
      (GPA_COUNTER                    ))
GO
  CREATE TABLE  LAB
    (LAB_TITLE                      char(15)            NOT NULL,
     LAB_DESCRIBTION                char(25)            NULL,
     LAB_CODE                       int                 NOT NULL,
     FK_COURSECOU_CODE              char(8)             NULL,
    CONSTRAINT labcode
    PRIMARY KEY NONCLUSTERED
      (LAB_CODE                       ))
GO
  CREATE TABLE  MAJOR
    (MAJ_TITLE                      char(30)            NOT NULL,
     MAJ_SNO                        int                 NOT NULL,
     MAJ_MIN_HIGH_SCHOOL_PERCENT    float               NULL,
     FK_DEPARTMENTDEP_ID            int                 NULL,
    CONSTRAINT majsno
    PRIMARY KEY NONCLUSTERED
      (MAJ_SNO                        ))
GO
  CREATE TABLE  MARK
    (MAR_COUNTER                    int                 NOT NULL,
     MAR_MARK                       int                 NOT NULL,
     MAR_DATE                       datetime            NULL,
     FK_STUDENTSTU_REGISTRATION_NUM float               NULL,
     FK_EXAMEXA_NUMBER              int                 NULL,
    CONSTRAINT marcounter
    PRIMARY KEY NONCLUSTERED
      (MAR_COUNTER                    ))
GO
  CREATE TABLE  NATIONALITY
    (NAT_IDENTIFICATION             int                 NOT NULL,
     NAT_NATIONALITY                char(25)
       DEFAULT 'else'
```

```
      NOT NULL ,
    CONSTRAINT natidentification
    PRIMARY KEY NONCLUSTERED
      (NAT_IDENTIFICATION                    ))
GO
  CREATE TABLE  PAYMENT
    (PAY_RECEIPT_NO                          float                  NOT NULL,
     PAY_DATE                                datetime               NULL,
     PAY_METHOD                              char(10)
       DEFAULT 'cash'
       NULL ,
     PAY_CURRENCY                            char(15)
       DEFAULT 'egyptian pound'
       NULL ,
     PAY_DISCOUNTED                          char(1)
       DEFAULT 'n'
       NULL ,
     FK_STUDENTSTU_REGISTRATION_NUM          float                  NULL,
     FK_TUITIONTUI_COUNTER                   int                    NULL,
    CONSTRAINT payreceiptno
    PRIMARY KEY NONCLUSTERED
      (PAY_RECEIPT_NO                        ))
GO
  CREATE TABLE  PENALTY
    (PEN_NAME                                char(15)               NOT NULL,
     PEN_SERIAL                              int                    NOT NULL,
     PEN_CONCEQUENCES                        char(35)               NULL,
    CONSTRAINT penserial
    PRIMARY KEY NONCLUSTERED
      (PEN_SERIAL                            ))
GO
  CREATE TABLE  PENALTY_STUDENT
    (PEN_STU_SNO                             int                    NOT NULL,
     PEN_STU_DATE                            datetime               NULL,
     FK_STUDENTSTU_REGISTRATION_NUM          float                  NULL,
     FK_PENALTYPEN_SERIAL                    int                    NULL,
    CONSTRAINT penstusno
    PRIMARY KEY NONCLUSTERED
      (PEN_STU_SNO                           ))
GO
  CREATE TABLE  PREREQUISITE
    (PRE_SNO                                 int                    NOT NULL,
     PRE_DETAIL1                             char(8)                NOT NULL,
     PRE_DETAIL2                             char(8)                NULL,
     PRE_DETAIL3                             char(8)                NULL,
     PRE_DETAIL4                             char(8)                NULL,
     PRE_DETAIL5                             char(8)                NULL,
     FK_COURSECOU_CODE                       char(8)                NULL,
    CONSTRAINT presno
    PRIMARY KEY NONCLUSTERED
      (PRE_SNO                               ))
GO
  CREATE TABLE  REGISTRATION
    (REG_SERIALNO                            int                    NOT NULL,
     REG_DATE                                datetime               NULL,
     FK_COURSECOU_CODE                       char(8)                NULL,
     FK_PAYMENTPAY_RECEIPT_NO                float                  NULL,
     FK_STUDENTSTU_REGISTRATION_NUM          float                  NULL,
    CONSTRAINT regserialno
    PRIMARY KEY NONCLUSTERED
      (REG_SERIALNO                          ))
```

```
GO
  CREATE TABLE  SEMESTER
    (SEM_NAME                          char(12)
       DEFAULT 'September'
       NOT NULL ,
     SEM_YEAR                          int                    NOT NULL,
     SEM_CODE                          int                    NOT NULL,
    CONSTRAINT semcode
    PRIMARY KEY NONCLUSTERED
      (SEM_CODE                         ))
GO
  CREATE TABLE  STUDENT
    (STU_REGISTRATION_NUMBER           float                  NOT NULL,
     FK_APPLICANTAPP_CODE              int                    NULL,
     FK_AUTHORITYAUT_CODE              int                    NULL,
    CONSTRAINT sturegistrationnumber
    PRIMARY KEY NONCLUSTERED
      (STU_REGISTRATION_NUMBER          ))
GO
  CREATE TABLE  STUDENT_ASSISTANTSHIP
    (STU_ASS_SERIAL_NO                 int                    NOT NULL,
     STU_ASS_STARTS                    datetime               NOT NULL,
     STU_ASS_ENDS                      datetime               NOT NULL,
     FK_SEMESTERSEM_CODE               int                    NULL,
     FK_ASSISTANTSHIASS_NUMBER         int                    NULL,
     FK_STUDENTSTU_REGISTRATION_NUM    float                  NULL,
    CONSTRAINT stuassserialno
    PRIMARY KEY NONCLUSTERED
      (STU_ASS_SERIAL_NO                ))
GO
  CREATE TABLE  TUITION
    (TUI_COUNTER                       int                    NOT NULL,
     TUI_AMOUNT                        float                  NOT NULL,
     TUI_CURRENCY                      char(15)
       DEFAULT 'egyptian pound'
       NOT NULL ,
     FK_SEMESTERSEM_CODE               int                    NULL,
     FK_MAJORMAJ_SNO                   int                    NULL,
    CONSTRAINT tuicounter
    PRIMARY KEY NONCLUSTERED
      (TUI_COUNTER                      ))
GO
  CREATE TABLE  UNIVERSITY
    (UNI_NAME                          char(35)               NOT NULL,
     UNI_ID                            int                    NOT NULL,
     UNI_COUNTRY                       char(15)               NULL,
    CONSTRAINT uniid
    PRIMARY KEY NONCLUSTERED
      (UNI_ID                           ))
GO
  CREATE NONCLUSTERED INDEX  I0000053
    ON  APPLICANT
    (FK_UNIVERSITYUNI_ID                )
GO
  CREATE NONCLUSTERED INDEX  I0000055
    ON  PENALTY_STUDENT
    (FK_STUDENTSTU_REGISTRATION_NUM     )
GO
  CREATE NONCLUSTERED INDEX  I0000057
    ON  PENALTY_STUDENT
    (FK_PENALTYPEN_SERIAL               )
```

```
GO
  CREATE NONCLUSTERED INDEX  I0000059
    ON  COURSE_MAJOR
    (FK_COURSECOU_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000061
    ON  COURSE_MAJOR
    (FK_MAJORMAJ_SNO                      )
GO
  CREATE NONCLUSTERED INDEX  I0000063
    ON  STUDENT_ASSISTANTSHIP
    (FK_SEMESTERSEM_CODE                  )
GO
  CREATE NONCLUSTERED INDEX  I0000065
    ON  STUDENT_ASSISTANTSHIP
    (FK_ASSISTANTSHIASS_NUMBER            )
GO
  CREATE NONCLUSTERED INDEX  I0000067
    ON  STUDENT_ASSISTANTSHIP
    (FK_STUDENTSTU_REGISTRATION_NUM       )
GO
  CREATE NONCLUSTERED INDEX  I0000069
    ON  LAB
    (FK_COURSECOU_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000071
    ON  PREREQUISITE
    (FK_COURSECOU_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000073
    ON  EXAM
    (FK_COURSECOU_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000075
    ON  REGISTRATION
    (FK_COURSECOU_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000077
    ON  GPA
    (FK_STUDENTSTU_REGISTRATION_NUM       )
GO
  CREATE NONCLUSTERED INDEX  I0000079
    ON  MARK
    (FK_STUDENTSTU_REGISTRATION_NUM       )
GO
  CREATE NONCLUSTERED INDEX  I0000081
    ON  MARK
    (FK_EXAMEXA_NUMBER                    )
GO
  CREATE NONCLUSTERED INDEX  I0000083
    ON  PAYMENT
    (FK_STUDENTSTU_REGISTRATION_NUM       )
GO
  CREATE NONCLUSTERED INDEX  I0000085
    ON  REGISTRATION
    (FK_PAYMENTPAY_RECEIPT_NO             )
GO
  CREATE NONCLUSTERED INDEX  I0000087
    ON  PAYMENT
    (FK_TUITIONTUI_COUNTER                )
GO
```

```
  CREATE NONCLUSTERED INDEX  I0000089
    ON  TUITION
    (FK_SEMESTERSEM_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000091
    ON  TUITION
    (FK_MAJORMAJ_SNO                        )
GO
  CREATE NONCLUSTERED INDEX  I0000093
    ON  APPLICANT
    (FK_BATCHBAT_NUMBER                     )
GO
  CREATE NONCLUSTERED INDEX  I0000095
    ON  STUDENT
    (FK_APPLICANTAPP_CODE                   )
GO
  CREATE NONCLUSTERED INDEX  I0000097
    ON  APPLICANT
    (FK_MAJORMAJ_SNO                        )
GO
  CREATE NONCLUSTERED INDEX  I0000099
    ON  APPLICANT
    (FK_CERTIFICATECER_NUMBER               )
GO
  CREATE NONCLUSTERED INDEX  I0000101
    ON  APPLICANT
    (FK_NATIONALITYNAT_IDENTIFICATI         )
GO
  CREATE NONCLUSTERED INDEX  I0000103
    ON  REGISTRATION
    (FK_STUDENTSTU_REGISTRATION_NUM         )
GO
  CREATE NONCLUSTERED INDEX  I0000105
    ON  STUDENT
    (FK_AUTHORITYAUT_CODE                   )
GO
  CREATE NONCLUSTERED INDEX  I0000107
    ON  BATCH
    (FK_SEMESTERSEM_CODE                    )
GO
  CREATE NONCLUSTERED INDEX  I0000109
    ON  DEPARTMENT
    (FK_COLLEGECOL_SERIAL                   )
GO
  CREATE NONCLUSTERED INDEX  I0000111
    ON  MAJOR
    (FK_DEPARTMENTDEP_ID                    )
GO
  ALTER TABLE APPLICANT
    ADD
    FOREIGN KEY
      (FK_NATIONALITYNAT_IDENTIFICATI       )
     REFERENCES NATIONALITY
GO
  ALTER TABLE APPLICANT
    ADD
    FOREIGN KEY
      (FK_CERTIFICATECER_NUMBER             )
      REFERENCES CERTIFICATE
GO
  ALTER TABLE APPLICANT
```

```
    ADD
    FOREIGN KEY
      (FK_MAJORMAJ_SNO                    )
     REFERENCES MAJOR
GO
  ALTER TABLE APPLICANT
    ADD
    FOREIGN KEY
      (FK_BATCHBAT_NUMBER                 )
      REFERENCES BATCH
GO
  ALTER TABLE APPLICANT
    ADD
    FOREIGN KEY
      (FK_UNIVERSITYUNI_ID                )
      REFERENCES UNIVERSITY
GO
  ALTER TABLE BATCH
    ADD
    FOREIGN KEY
      (FK_SEMESTERSEM_CODE                )
      REFERENCES SEMESTER
GO
  ALTER TABLE COURSE_MAJOR
    ADD
    FOREIGN KEY
      (FK_MAJORMAJ_SNO                    )
      REFERENCES MAJOR
GO
  ALTER TABLE COURSE_MAJOR
    ADD
    FOREIGN KEY
      (FK_COURSECOU_CODE                  )
      REFERENCES COURSE
GO
  ALTER TABLE DEPARTMENT
    ADD
    FOREIGN KEY
      (FK_COLLEGECOL_SERIAL               )
      REFERENCES COLLEGE
GO
  ALTER TABLE EXAM
    ADD
    FOREIGN KEY
      (FK_COURSECOU_CODE                  )
      REFERENCES COURSE
GO
  ALTER TABLE GPA
    ADD
    FOREIGN KEY
      (FK_STUDENTSTU_REGISTRATION_NUM     )
      REFERENCES STUDENT
GO
  ALTER TABLE LAB
    ADD
    FOREIGN KEY
      (FK_COURSECOU_CODE                  )
      REFERENCES COURSE
GO
  ALTER TABLE MAJOR
    ADD
```

```
    FOREIGN KEY
      (FK_DEPARTMENTDEP_ID                )
      REFERENCES DEPARTMENT
GO
  ALTER TABLE MARK
    ADD
    FOREIGN KEY
      (FK_EXAMEXA_NUMBER                  )
      REFERENCES EXAM
GO
  ALTER TABLE MARK
    ADD
    FOREIGN KEY
      (FK_STUDENTSTU_REGISTRATION_NUM     )
      REFERENCES STUDENT
GO
  ALTER TABLE PAYMENT
    ADD
    FOREIGN KEY
      (FK_STUDENTSTU_REGISTRATION_NUM     )
      REFERENCES STUDENT
GO
  ALTER TABLE PENALTY_STUDENT
    ADD
    FOREIGN KEY
      (FK_PENALTYPEN_SERIAL               )
      REFERENCES PENALTY
GO
  ALTER TABLE PENALTY_STUDENT
    ADD
    FOREIGN KEY
      (FK_STUDENTSTU_REGISTRATION_NUM     )
      REFERENCES STUDENT
GO
  ALTER TABLE PREREQUISITE
    ADD
    FOREIGN KEY
      (FK_COURSECOU_CODE                  )
      REFERENCES COURSE
GO

  ALTER TABLE REGISTRATION
    ADD
    FOREIGN KEY
      (FK_STUDENTSTU_REGISTRATION_NUM     )
      REFERENCES STUDENT
GO
  ALTER TABLE REGISTRATION
    ADD
    FOREIGN KEY
      (FK_PAYMENTPAY_RECEIPT_NO           )
      REFERENCES PAYMENT
GO
  ALTER TABLE REGISTRATION
    ADD
    FOREIGN KEY
      (FK_COURSECOU_CODE                  )
      REFERENCES COURSE
GO
  ALTER TABLE STUDENT
    ADD
```

```
    FOREIGN KEY
      (FK_APPLICANTAPP_CODE                )
      REFERENCES APPLICANT
GO
  ALTER TABLE STUDENT
    ADD
    FOREIGN KEY
      (FK_AUTHORITYAUT_CODE                )
      REFERENCES AUTHORITY
GO
  ALTER TABLE STUDENT_ASSISTANTSHIP
    ADD
    FOREIGN KEY
      (FK_ASSISTANTSHIASS_NUMBER           )
      REFERENCES ASSISTANTSHIP
GO
  ALTER TABLE STUDENT_ASSISTANTSHIP
    ADD
    FOREIGN KEY
      (FK_SEMESTERSEM_CODE                 )
      REFERENCES SEMESTER
GO
  ALTER TABLE STUDENT_ASSISTANTSHIP
    ADD
    FOREIGN KEY
      (FK_STUDENTSTU_REGISTRATION_NUM      )
      REFERENCES STUDENT
GO
  ALTER TABLE TUITION
    ADD
    FOREIGN KEY
      (FK_SEMESTERSEM_CODE                 )
      REFERENCES SEMESTER
GO
  ALTER TABLE TUITION
    ADD
    FOREIGN KEY
      (FK_MAJORMAJ_SNO                     )
      REFERENCES MAJOR
GO
  /*************************************************************/
  /*   Trigger of PAYMENT_INS is to enforce Foreign Keys */
  /*   integrity, when inserting table PAYMENT. */
  /*************************************************************/
  CREATE TRIGGER PAYMENT_INS
    ON PAYMENT
    FOR INSERT AS
    DECLARE
      @row INT,
      @nullrow INT
    SELECT @row = @@rowcount
  BEGIN
    /* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++ */
    /*   Can not modify or create Foreign Keys of table PAYMENT, */
    /*   when no matching Primary key of table TUITION exists. */
    /* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++ */

    IF ((SELECT DISTINCT inserted.FK_TUITIONTUI_COUNTER FROM inserted) is
not null    )
    BEGIN
      SELECT @nullrow = (select count (*) from inserted
```

```
      WHERE
      FK_TUITIONTUI_COUNTER = null )
      IF (SELECT COUNT(*) FROM inserted, TUITION
      WHERE
      TUITION.TUI_COUNTER                          =
inserted.FK_TUITIONTUI_COUNTER
      ) != @row - @nullrow
      BEGIN
      raiserror 20053 "INVALID FOREIGN KEY VALUE OF (PAYMENT), PRIMARY KEY
OF (TUITION) NOT FOUND"
      ROLLBACK TRANSACTION
      END
    END
  END
GO
  /*************************************************************/
  /*  Trigger of PAYMENT_UPD is to enforce Foreign and Primary */
  /*  Keys integrity, when updating table PAYMENT. */
  /*************************************************************/
  CREATE TRIGGER PAYMENT_UPD
    ON PAYMENT
    FOR UPDATE AS
    DECLARE
      @row INT,
      @nullrow INT
    SELECT @row = @@rowcount
  BEGIN
  IF UPDATE(FK_TUITIONTUI_COUNTER)
  BEGIN
    /* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++ */
    /*   Can not modify or create Foreign Keys of table PAYMENT, */
    /*   when no matching Primary key of table TUITION exists. */
    /* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++ */

    IF ((SELECT DISTINCT inserted.FK_TUITIONTUI_COUNTER FROM inserted) is
not null    )
    BEGIN
      SELECT @nullrow = (select count (*) from inserted
      WHERE
      FK_TUITIONTUI_COUNTER = null )
      IF (SELECT COUNT(*) FROM inserted, TUITION
      WHERE
      TUITION.TUI_COUNTER                          =
inserted.FK_TUITIONTUI_COUNTER
      ) != @row - @nullrow
      BEGIN
      raiserror 20053 "INVALID FOREIGN KEY VALUE OF (PAYMENT), PRIMARY KEY
OF (TUITION) NOT FOUND"
      ROLLBACK TRANSACTION
      END
    END
  END
  END
GO
  /*************************************************************/
  /*   Trigger of TUITION_UPD is to enforce Foreign and  */
  /*   Primary Keys integrity, when updating table TUITION. */
  /*************************************************************/

  CREATE TRIGGER TUITION_UPD
    ON TUITION
```

```
    FOR UPDATE AS
    DECLARE
      @row INT,
      @nullrow INT
    SELECT @row = @@rowcount
  BEGIN
    -- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
    --   Can not modify Primary Key of table TUITION,
    --   when Foreign key of table PAYMENT exists.
    -- ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
    IF EXISTS(SELECT 1
      FROM PAYMENT, inserted, deleted
    WHERE
      PAYMENT.FK_TUITIONTUI_COUNTER = deleted.TUI_COUNTER
        AND (
      deleted.TUI_COUNTER != inserted.TUI_COUNTER ))
      BEGIN
      raiserror 20052 "Foreign Key of (TUITION) exists, Primary Key of
(PAYMENT) can not be updated "
      ROLLBACK TRANSACTION
      END
  END
GO

  /*************************************************************/
  /*  Trigger of TUITION_DEL to Set Null or Cascade */
  /*  delete Foreign Key rows, when deleting Primary Key row */
  /*  of table TUITION. */
  /*************************************************************/

  CREATE TRIGGER TUITION_DEL
    ON TUITION
    FOR DELETE AS
    /* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++ */
    /*   Set Null Foreign Key rows of table PAYMENT, */
    /*   when deleting Primary Key row of table TUITION. */
    /* ++++++++++++++++++++++++++++++++++++++++++++++++++++++++ */
    UPDATE PAYMENT
    SET
      PAYMENT.FK_TUITIONTUI_COUNTER = NULL
    FROM deleted, PAYMENT
    WHERE
      PAYMENT.FK_TUITIONTUI_COUNTER                = deleted.TUI_COUNTER
GO
```

3-Data Model List

```
Type          Name
Subject Area   +-AR_DSS
Subject Area   | +-APPLICANTS
Entity Type    | | +-APPLICANT
Attribute      | | |   /APP_AGE                              (Number, 3,
Optional, Derived)
Attribute      | | |   /APP_PREDICTED_GRADUATION_GRADE_M (Text, 20,
Optional, Derived)
Attribute      | | |   /APP_PREDICTED_MAJOR_F            (Text, 30,
Optional, Derived)
Attribute      | | |   APP_FULL_NAME                     (Text, 35,
Mandatory, Basic)
Attribute      | | |   APP_DOB                           (Date, 8,
Mandatory, Basic)
Attribute      | | |   APP_GENDER                        (Text, 1,
Optional, Basic)
Attribute      | | |   APP_ADDRESS                       (Text, 35,
Optional, Basic)
Attribute      | | |   APP_PREVIOUSLY_ABANDONED          (Text, 1,
Optional, Basic)
Attribute      | | |   APP_TRANSFERRED                   (Text, 1,
Optional, Basic)
Attribute      | | | I APP_CODE                          (Number, 6,
Mandatory, Basic)
Attribute      | | |   APP_CERTIFICATE_YEAR              (Number, 5,
Mandatory, Basic)
Attribute      | | |   APP_CERTIFICATION_PERCENT         (Number, 4,
Mandatory, Basic)
Attribute      | | |   APP_TELEPHONE                     (Text, 15,
Optional, Basic)
Attribute      | | |   APP_INTERVIEWED                   (Text, 1,
Optional, Basic)
Attribute      | | |   /APP_HOLD_TO_NEXT_BATCH_G         (Text, 1,
Optional, Derived)
Attribute      | | |   /APP_ACCEPTED_A                   (Text, 1,
Optional, Derived)
Relationship   | | |   Sometimes TRANSFERRED_FROM One UNIVERSITY
Relationship   | | |   Always POINTS_TO One BATCH
Relationship   | | |   Sometimes BECOMES One STUDENT
Relationship   | | |   Always HAS One MAJOR
Relationship   | | |   Always REFERS_TO One CERTIFICATE
Relationship   | | |   Sometimes HAS One NATIONALITY
               | | +-
               | +-
Subject Area   | +-BATCHES
Entity Type    | | +-BATCH
Attribute      | | |   BAT_TITLE                  (Text, 10, Optional,
Basic)
Attribute      | | | I BAT_NUMBER                 (Number, 5, Mandatory,
Basic)
Attribute      | | |   BAT_CEILING                (Number, 5, Mandatory,
Basic)
Attribute      | | |   BAT_YEAR                   (Number, 5, Mandatory,
Basic)
Attribute      | | |   BAT_OPEN_DATE              (Date, 8, Optional,
Basic)
Attribute      | | |   BAT_CLOSING_DATE           (Date, 8, Mandatory,
Basic)
```

```
Attribute       | | |   BAT_MARKET_SHARE              (Number, 5, Optional,
Basic)
Attribute       | | |   BAT_GOVERNMENT_STAT           (Number, 6, Optional,
Basic)
Attribute       | | |   /BAT_PREDICTED_APPLICANTS_C (Number, 6, Mandatory,
Derived)
Relationship    | | |   Sometimes MAINTAINS One or More APPLICANT
Relationship    | | |   Always IS_INITIATED_BY One SEMESTER
                | | +-
Entity Type     | | +-SEMESTER
Attribute       | | |   SEM_NAME (Text, 12, Mandatory, Basic)
Attribute       | | |   SEM_YEAR (Number, 5, Mandatory, Basic)
Attribute       | | | I SEM_CODE (Number, 5, Mandatory, Basic)
Relationship    | | |   Sometimes TRIGGERS One or More TUITION
Relationship    | | |   Sometimes DEVELOPS One or More STUDENT_ASSISTANTSHIP
Relationship    | | |   Sometimes INITIATES One BATCH
                | | +-
                | +-
Subject Area    | +-BOOKINGS
Entity Type     | | +-REGISTRATION
Attribute       | | | I REG_SERIALNO (Number, 6, Mandatory, Basic)
Attribute       | | |   REG_DATE (Date, 8, Optional, Basic)
Relationship    | | |   Sometimes REQUIRES One PAYMENT
Relationship    | | |   Always IS_ESTABLISHED_BY One COURSE
Relationship    | | |   Always IS_DONE_BY One STUDENT
                | | +-
                | +-
Subject Area    | +-COURSES
Entity Type     | | +-COURSE
Attribute       | | |   COU_STAGE (Text, 15, Optional, Basic)
Attribute       | | |   COU_TITLE (Text, 25, Mandatory, Basic)
Attribute       | | | I COU_CODE (Text, 8, Mandatory, Basic)
Attribute       | | |   COU_CREDIT_HOURS (Number, 3, Mandatory, Basic)
Attribute       | | |   COU_PASS_MARK (Number, 5, Mandatory, Basic)
Attribute       | | |   COU_FULL_MARK (Number, 5, Mandatory, Basic)
Attribute       | | |   COU_AREA (Text, 25, Optional, Basic)
Relationship    | | |   Sometimes ESTABLISHES One or More EXAM
Relationship    | | |   Sometimes DEPENDS_ON One or More PREREQUISITE
Relationship    | | |   Sometimes REQUIRES One or More LAB
Relationship    | | |   Sometimes ESTABLISHES One or More REGISTRATION
Relationship    | | |   Sometimes POINTS_TO One or More COURSE_MAJOR
                | | +-
Entity Type     | | +-COURSE_MAJOR
Attribute       | | | I COU_MAJ_COUNTER (Number, 5, Mandatory, Basic)
Relationship    | | |   Always REFERS_TO One COURSE
Relationship    | | |   Always REFERS_TO One MAJOR
                | | +-
Entity Type     | | +-LAB
Attribute       | | |   LAB_TITLE (Text, 15, Mandatory, Basic)
Attribute       | | |   LAB_DESCRIBTION (Text, 25, Optional, Basic)
Attribute       | | | I LAB_CODE (Number, 5, Mandatory, Basic)
Relationship    | | |   Always NEEDS One COURSE
                | | +-
Entity Type     | | +-PREREQUISITE
Attribute       | | | I PRE_SNO (Number, 5, Mandatory, Basic)
Attribute       | | |   PRE_DETAIL1 (Text, 8, Mandatory, Basic)
Attribute       | | |   PRE_DETAIL2 (Text, 8, Optional, Basic)
Attribute       | | |   PRE_DETAIL3 (Text, 8, Optional, Basic)
Attribute       | | |   PRE_DETAIL4 (Text, 8, Optional, Basic)
Attribute       | | |   PRE_DETAIL5 (Text, 8, Optional, Basic)
Relationship    | | |   Always REFERS_TO One COURSE
```

```
                | | +-
                | +-
Subject Area    | +-EXAMINATIONS
Entity Type     | | +-EXAM
Attribute       | | | I EXA_NUMBER (Number, 5, Mandatory, Basic)
Attribute       | | |   EXA_TYPE (Text, 15, Mandatory, Basic)
Relationship    | | |   Always IS_FOR One COURSE
Relationship    | | |   Sometimes INITIATES One or More MARK
                | | +-
                | +-
Subject Area    | +-EXTERNAL_UNIVERSITIES
Entity Type     | | +-UNIVERSITY
Attribute       | | |   UNI_NAME (Text, 35, Mandatory, Basic)
Attribute       | | | I UNI_ID (Number, 5, Mandatory, Basic)
Attribute       | | |   UNI_COUNTRY (Text, 15, Optional, Basic)
Relationship    | | |   Sometimes SENDS One or More APPLICANT
                | | +-
                | +-
Subject Area    | +-FEES
Entity Type     | | +-PAYMENT
Attribute       | | | I PAY_RECEIPT_NO  (Number, 10, Mandatory, Basic)
Attribute       | | |   PAY_DATE        (Date, 8, Optional, Basic)
Attribute       | | |   PAY_METHOD      (Text, 10, Optional, Basic)
Attribute       | | |   /PAY_AMOUNT     (Number, 7, Mandatory, Derived)
Attribute       | | |   PAY_CURRENCY    (Text, 15, Optional, Basic)
Attribute       | | |   PAY_DISCOUNTED  (Text, 1, Optional, Basic)
Attribute       | | |   /PAY_DISCOUNT   (Number, 4, Optional, Derived)
Attribute       | | |   /PAY_NET_AMOUNT (Number, 6, Optional, Derived)
Relationship    | | |   Always IS_FOR One STUDENT
Relationship    | | |   Sometimes FULFILLS One or More REGISTRATION
Relationship    | | |   Sometimes IS_BASED_ON One TUITION
                | | +-
Entity Type     | | +-TUITION
Attribute       | | | I TUI_COUNTER (Number, 6, Mandatory, Basic)
Attribute       | | |   TUI_AMOUNT (Number, 7, Mandatory, Basic)
Attribute       | | |   TUI_CURRENCY (Text, 15, Mandatory, Basic)
Relationship    | | |   Sometimes IS_BASED_ON One SEMESTER
Relationship    | | |   Always POINTS_TO One MAJOR
Relationship    | | |   Sometimes HAS One or More PAYMENT
                | | +-
                | +-
Subject Area    | +-HIGH_SCHOOLS
Entity Type     | | +-CERTIFICATE
Attribute       | | | I CER_NUMBER (Number, 5, Mandatory, Basic)
Attribute       | | |   CER_NAME (Text, 35, Mandatory, Basic)
Attribute       | | |   CER_ORIGIN (Text, 25, Optional, Basic)
Relationship    | | |   Sometimes POINTS_TO One or More APPLICANT
                | | +-
                | +-
Subject Area    | +-INSTITUTES
Entity Type     | | +-COLLEGE
Attribute       | | |   COL_NAME (Text, 35, Mandatory, Basic)
Attribute       | | | I COL_SERIAL (Number, 5, Mandatory, Basic)
Attribute       | | |   COL_LOCATION (Text, 25, Optional, Basic)
Relationship    | | |   Sometimes MAINTAINS One or More DEPARTMENT
                | | +-
Entity Type     | | +-DEPARTMENT
Attribute       | | |   DEP_TITLE (Text, 20, Mandatory, Basic)
Attribute       | | | I DEP_ID (Number, 5, Mandatory, Basic)
Attribute       | | |   DEP_LOCATION (Text, 25, Optional, Basic)
Attribute       | | |   DEP_TYPE (Text, 15, Optional, Basic)
```

```
Relationship   | | |   Sometimes OFFERS One or More MAJOR
Relationship   | | |   Always IS_MAINTAINED_BY One COLLEGE
               | | +-
Entity Type    | | +-MAJOR
Attribute      | | |   MAJ_TITLE (Text, 30, Mandatory, Basic)
Attribute      | | | I MAJ_SNO (Number, 5, Mandatory, Basic)
Attribute      | | |   MAJ_MIN_HIGH_SCHOOL_PERCENT (Number, 4, Optional,
Basic)
Relationship   | | |   Sometimes DEVELOPS One or More TUITION
Relationship   | | |   Sometimes OFFERS One or More COURSE_MAJOR
Relationship   | | |   Sometimes HAS One or More APPLICANT
Relationship   | | |   Always IS_OFFERED_BY One DEPARTMENT
               | | +-
               | +-
Subject Area   | +-NATIONALITIES
Entity Type    | | +-NATIONALITY
Attribute      | | | I NAT_IDENTIFICATION (Number, 5, Mandatory, Basic)
Attribute      | | |   NAT_NATIONALITY (Text, 25, Mandatory, Basic)
Relationship   | | |   Sometimes HAS One or More APPLICANT
               | | +-
               | +-
Subject Area   | +-PUNISHMENTS
Entity Type    | | +-PENALTY
Attribute      | | |   PEN_NAME (Text, 15, Mandatory, Basic)
Attribute      | | | I PEN_SERIAL (Number, 5, Mandatory, Basic)
Attribute      | | |   PEN_CONCEQUENCES (Text, 35, Optional, Basic)
Relationship   | | |   Sometimes DEVELOPS One or More PENALTY_STUDENT
               | | +-
Entity Type    | | +-PENALTY_STUDENT
Attribute      | | | I PEN_STU_SNO (Number, 5, Mandatory, Basic)
Attribute      | | |   PEN_STU_DATE (Date, 8, Optional, Basic)
Relationship   | | |   Always POINTS_TO One STUDENT
Relationship   | | |   Always REFERS_TO One PENALTY
               | | +-
               | +-
Subject Area   | +-RECORDS
Entity Type    | | +-GPA
Attribute      | | | I GPA_COUNTER (Number, 6, Mandatory, Basic)
Relationship   | | |   Always REFERS_TO One STUDENT
               | | +-
               | +-
Subject Area   | +-RESULTS
Entity Type    | | +-MARK
Attribute      | | | I MAR_COUNTER (Number, 6, Mandatory, Basic)
Attribute      | | |   MAR_MARK    (Number, 5, Mandatory, Basic)
Attribute      | | |   MAR_DATE    (Date, 8, Optional, Basic)
Attribute      | | |   /MAR_POINTS (Number, 2, Mandatory, Derived)
Relationship   | | |   Always FOR One STUDENT
Relationship   | | |   Always POINTS_TO One EXAM
               | | +-
               | +-
Subject Area   | +-SCHOLARSHIPS
Entity Type    | | +-ASSISTANTSHIP
Attribute      | | |   ASS_TITLE (Text, 30, Mandatory, Basic)
Attribute      | | | I ASS_NUMBER (Number, 5, Mandatory, Basic)
Attribute      | | |   ASS_REQUIREMENTS (Text, 30, Optional, Basic)
Attribute      | | |   ASS_DISCOUNT_RATE (Number, 4, Mandatory, Basic)
Attribute      | | |   ASS_CATEGORY (Text, 25, Optional, Basic)
Relationship   | | |   Sometimes POINTS_TO One or More
STUDENT_ASSISTANTSHIP
               | | +-
```

```
                | +-
Subject Area    | +-SPONSORS
Entity Type     | | +-AUTHORITY
Attribute       | | |   AUT_NAME (Text, 25, Mandatory, Basic)
Attribute       | | | I AUT_CODE (Number, 5, Mandatory, Basic)
Attribute       | | |   AUT_ADDRESS (Text, 35, Optional, Basic)
Attribute       | | |   AUT_TELEPHONE (Text, 12, Optional, Basic)
Attribute       | | |   AUT_CONTACT_PERSON (Text, 25, Optional, Basic)
Relationship    | | |   Sometimes PAYS_FOR One or More STUDENT
                | | +-
                | +-
Subject Area    | +-STUDENTS
Entity Type     | | +-STUDENT
Attribute       | | |   /STU_TOTAL_COURSES            (Number, 2, Optional,
Derived)
Attribute       | | |   /STU_PREDICTED_PERFORMANCE_J  (Text, 20, Optional,
Derived)
Attribute       | | |   /STU_GPA                      (Number, 4, Optional,
Derived)
Attribute       | | | I STU_REGISTRATION_NUMBER       (Number, 10,
Mandatory, Basic)
Attribute       | | |   /STU_TOTAL_CREDIT_HOURS_REG   (Number, 4, Optional,
Derived)
Attribute       | | |   /STU_TOTAL_CREDIT_HOURS_ACH   (Number, 4, Optional,
Derived)
Attribute       | | |   /STU_GRADUATION_STATE         (Text, 1, Optional,
Derived)
Attribute       | | |   /STU_ABANDONMENT_STATE_R      (Text, 1, Optional,
Derived)
Attribute       | | |   /STU_ON_PROBATION_STATE_N     (Text, 1, Optional,
Derived)
Attribute       | | |   /STU_PREDICTED_ON_PROBATION_O (Text, 1, Optional,
Derived)
Relationship    | | |   Sometimes HAS One or More MARK
Relationship    | | |   Sometimes HAS One GPA
Relationship    | | |   Sometimes GOT One or More PENALTY_STUDENT
Relationship    | | |   Sometimes MAKES One or More PAYMENT
Relationship    | | |   Sometimes INITIATES One or More REGISTRATION
Relationship    | | |   Always WAS One APPLICANT
Relationship    | | |   Sometimes TAKES One or More STUDENT_ASSISTANTSHIP
Relationship    | | |   Sometimes IS_PAID_BY One AUTHORITY
                | | +-
Entity Type     | | +-STUDENT_ASSISTANTSHIP
Attribute       | | | I STU_ASS_SERIAL_NO (Number, 6, Mandatory, Basic)
Attribute       | | |   STU_ASS_STARTS (Date, 8, Mandatory, Basic)
Attribute       | | |   STU_ASS_ENDS (Date, 8, Mandatory, Basic)
Relationship    | | |   Always IS_GRANTED_ON One SEMESTER
Relationship    | | |   Always REFERS_TO One ASSISTANTSHIP
Relationship    | | |   Always REFERS_TO One STUDENT
                | | +-
                | +-
                +-
```

4- Matrices

4-1 Information needs/CSF

Cell Values:

= Not referenced

X = Include

1 2 3

4 5 6

7 8 9

Information Need	Crit Success Factor: MANAGING ADMISSION	Crit Success Factor: MANAGING REGISTRATION
CLUS M PREDICT STUDENT PERFORMAN		9
CLUS L CLASSIFY STUDENT TO GROUP		9
CLUS F APPLICANT MAJOR MATCH	9	
SQL P PERFORMANC AND DEPARTMENTS		9
SQL J PREDICT STUDENT PERFORMANC		9
SQL C APPLICANTS PREDICTION	9	
SQL R ABANDONMENT		9
SQL N SET TO ON PROBATION		9
SQL H ACCEPT REJECT TRANSFER	9	
SQL G HOLD APPLICANT	9	
SQL A ACCEPT REJECT	9	

4-2 Information needs/Entity types

Cell Values:
= Not referenced
X = Include

1	2	3
4	5	6
7	8	9

Entity Type	SQL A ACCEPT REJECT	SQL G HOLD APPLICANT	SQL H ACCEPT REJECT TRAN>	SQL N SET TO ON PROBATION	SQL R ABANDONMENT	SQL C APPLICANTS PREDICT>	SQL J PREDICT STUDENT PE>	SQL P PERFORMANC AND DEP>	CLUS F APPLICANT MAJOR M>	CLUS L CLASSIFY STUDENT >	CLUS M PREDICT STUDENT P>
EXAM											
DEPARTMENT								9			
MAJOR								9			
COLLEGE											
AUTHORITY											
NATIONALITY											
CERTIFICATE											
ASSISTANTSHIP											
SEMESTER											
BATCH						9					
REGISTRATION											
STUDENT				9	9		9	9			
APPLICANT	9	9	9						9	9	9
TUITION											
PAYMENT											
MARK											
GPA											
PENALTY											
COURSE											
PREREQUISITE											
LAB											
STUDENT ASSISTANTSHIP											
COURSE MAJOR											
PENALTY STUDENT											
UNIVERSITY											

4-3 Entity types/Elementary processes

Cell Values:
= Not referenced
C = Create
D = Delete
U = Update
R = Read only

Elementary Process	EXAM	DEPARTMENT	MAJOR	COLLEGE	AUTHORITY	NATIONALITY	CERTIFICATE	ASSISTANTSHIP	SEMESTER	BATCH	REGISTRATION	STUDENT	APPLICANT	TUITION	PAYMENT	MARK	GPA	PENALTY	COURSE	PREREQUISITE	LAB	STUDENT ASSISTANTSHIP	COURSE MAJOR	PENALTY STUDENT	UNIVERSITY
READ COLLEGE				R																					
DELETE COLLEGE		R		D																					
ADD DEPARTMENT		C		U																					
UPDATE DEPARTMENT		U																							
READ DEPARTMENT		R																							
DELETE DEPARTMENT		D	R																						
ADD MAJOR		U	C																						
UPDATE MAJOR			U																						
READ MAJOR			R																						
DELETE MAJOR			D										R	R									R		
ADD NATIONALITY						C																			
UPDATE NATIONALITY						U							U												
READ NATIONALITY						R																			
DELETE NATIONALITY						D							R												
ADD PENALTY																		C							
UPDATE PENALTY																		U						U	
READ PENALTY																		R							
DELETE PENALTY																		D						R	
ADD PENALTY STUDENT												U						U						C	
UPDATE PENALTY STUDENT																								U	
READ PENALTY STUDENT																								R	
DELETE PENALTY STUDENT																								D	
ADD MARK	U											U				C									
UPDATE MARK																U									
READ MARK																R									
DELETE MARK												R				D									
ADD ASSISTANTSHIP								C																	
UPDATE ASSISTANTSHIP								U																	
READ ASSISTANTSHIP								R																	
DELETE ASSISTANTSHIP								D														R			
ADD AUTHORITY					C																				
UPDATE AUTHORITY					U							U													
READ AUTHORITY					R																				
DELETE AUTHORITY					D							R													
ADD STUDENT												C	U												
UPDATE STUDENT												U													
READ STUDENT												R													
DELETE STUDENT					R						R	D			R	R	R					R		R	
ADD STUDENT ASSISTANTSHIP								U	U			U										C			
UPDATE STUDENT ASSISTANTSHIP																						U			
READ STUDENT ASSISTANTSHIP																						R			
DELETE STUDENT ASSISTANTSHIP																						D			
ADD GPA												U					C								
UPDATE GPA																	U								
READ GPA																	R								
DELETE GPA																	D								
ADD UNIVERSITY																									C
UPDATE UNIVERSITY													U												U
READ UNIVERSITY																									R
DELETE UNIVERSITY													R												D

Cell Values:

= Not referenced

C = Create

D = Delete

U = Update

R = Read only

Elementary Process	EXAM	DEPARTMENT	MAJOR	COLLEGE	AUTHORITY	NATIONALITY	CERTIFICATE	ASSISTANTSHIP	SEMESTER	BATCH	REGISTRATION	STUDENT	APPLICANT	TUITION	PAYMENT	MARK	GPA	PENALTY	COURSE	PREREQUISITE	LAB	STUDENT ASSISTANTSHIP	COURSE MAJOR	PENALTY STUDENT	UNIVERSITY
ADD PREREQUISITE																			U	C					
UPDATE PREREQUISITE																				U					
READ PREREQUISITE																				R					
DELETE PREREQUISITE																				D					
ADD APPLICANT			U			U	U			U			C												U
UPDATE APPLICANT													U												
READ APPLICANT													R												
DELETE APPLICANT												R	D												
ADD BATCH									U	C															
UPDATE BATCH										U															
READ BATCH										R															
DELETE BATCH										D			R												
ADD SEMESTER									C																
UPDATE SEMESTER									U					U											
READ SEMESTER									R																
DELETE SEMESTER									D	R				R								R			
ADD REGISTRATION											C	U			U				U						
READ REGISTRATION											R														
UPDATE REGISTRATION											U														
DELETE REGISTRATION											D	R							R						
ADD COURSE																			C						
UPDATE COURSE																			U						
READ COURSE																			R						
DELETE COURSE	R										R								D	R	R		R		
ADD COURSE MAJOR			U																U				C		
UPDATE COURSE MAJOR																							U		
READ COURSE MAJOR																							R		
DELETE COURSE MAJOR																							D		
ADD LAB																			U		C				
UPDATE LAB																					U				
READ LAB																					R				
DELETE LAB																					D				
ADD EXAM	C																		U						
UPDATE EXAM	U																								
READ EXAM	R																								
DELETE EXAM	D															R									
ADD PAYMENT												U		U	C										
UPDATE PAYMENT											U				U										
READ PAYMENT															R										
DELETE PAYMENT											R				D										
ADD TUITION			U						U					C											
UPDATE TUITION														U	U										
READ TUITION														R											
DELETE TUITION														D	R										
ADD CERTIFICATE							C																		
UPDATE CERTIFICATE							U						U												
READ CERTIFICATE							R																		
DELETE CERTIFICATE							D						R												
ADD COLLEGE				C																					
UPDATE COLLEGE				U																					

4-4 Business System/Elementary processes

Cell Values:
= Not referenced
X = Include

1 2 3
4 5 6
7 8 9

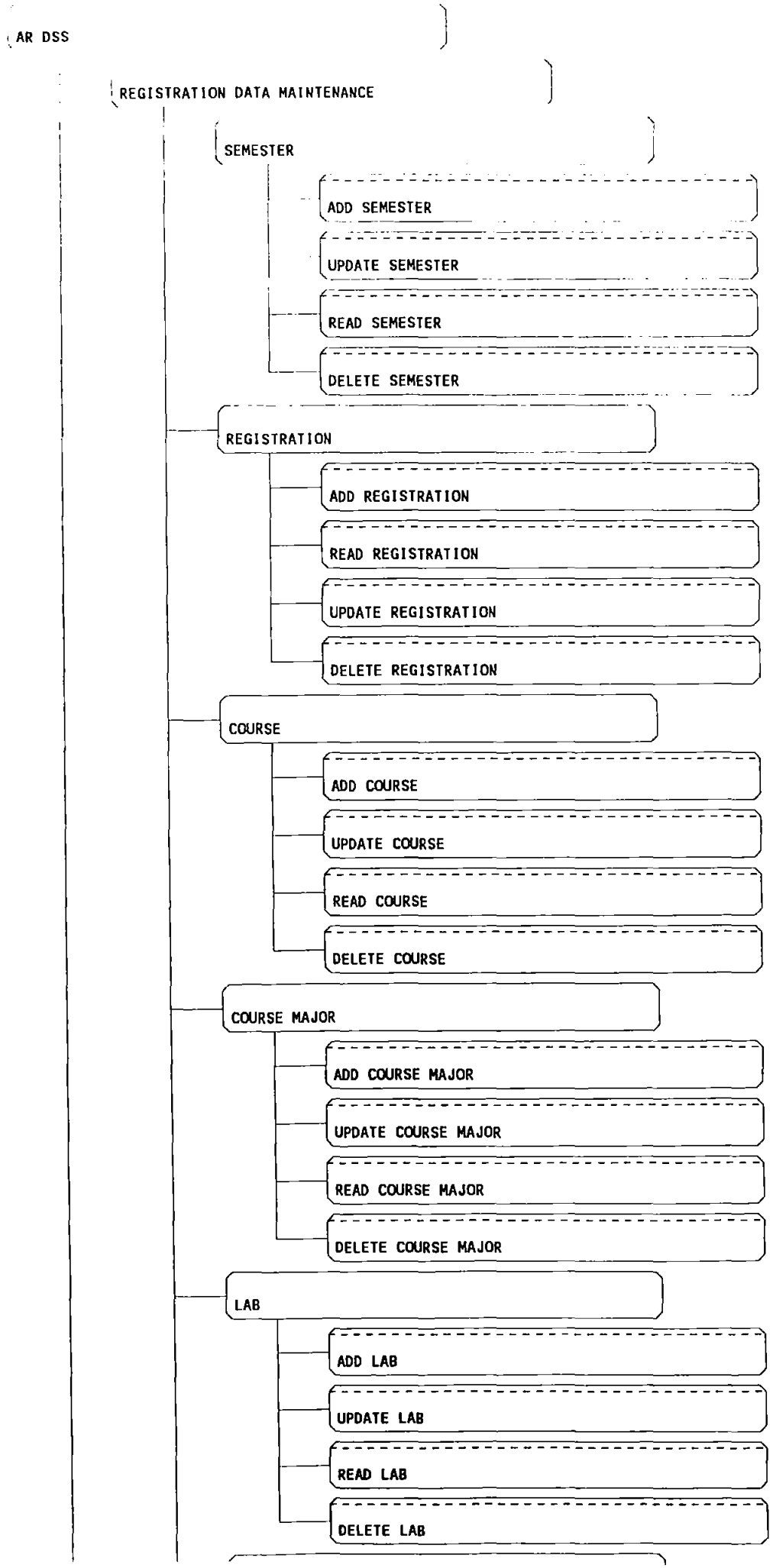
Elementary Process	Business System ADMISSION REGISTRATION >>
ADD APPLICANT	X
UPDATE APPLICANT	X
READ APPLICANT	X
DELETE APPLICANT	X
ADD BATCH	X
UPDATE BATCH	X
READ BATCH	X
DELETE BATCH	X
ADD SEMESTER	X
UPDATE SEMESTER	X
READ SEMESTER	X
DELETE SEMESTER	X
ADD REGISTRATION	X
READ REGISTRATION	X
UPDATE REGISTRATION	X
DELETE REGISTRATION	X
ADD COURSE	X
UPDATE COURSE	X
READ COURSE	X
DELETE COURSE	X
ADD COURSE MAJOR	X
UPDATE COURSE MAJOR	X
READ COURSE MAJOR	X
DELETE COURSE MAJOR	X
ADD LAB	X
UPDATE LAB	X
READ LAB	X
DELETE LAB	X
ADD EXAM	X
UPDATE EXAM	X
READ EXAM	X
DELETE EXAM	X
ADD PAYMENT	X
UPDATE PAYMENT	X
READ PAYMENT	X
DELETE PAYMENT	X
ADD TUITION	X
UPDATE TUITION	X
READ TUITION	X
DELETE TUITION	X
ADD CERTIFICATE	X
UPDATE CERTIFICATE	X
READ CERTIFICATE	X
DELETE CERTIFICATE	X
ADD COLLEGE	X
UPDATE COLLEGE	X
READ COLLEGE	X
DELETE COLLEGE	X
ADD DEPARTMENT	X
UPDATE DEPARTMENT	X
READ DEPARTMENT	X

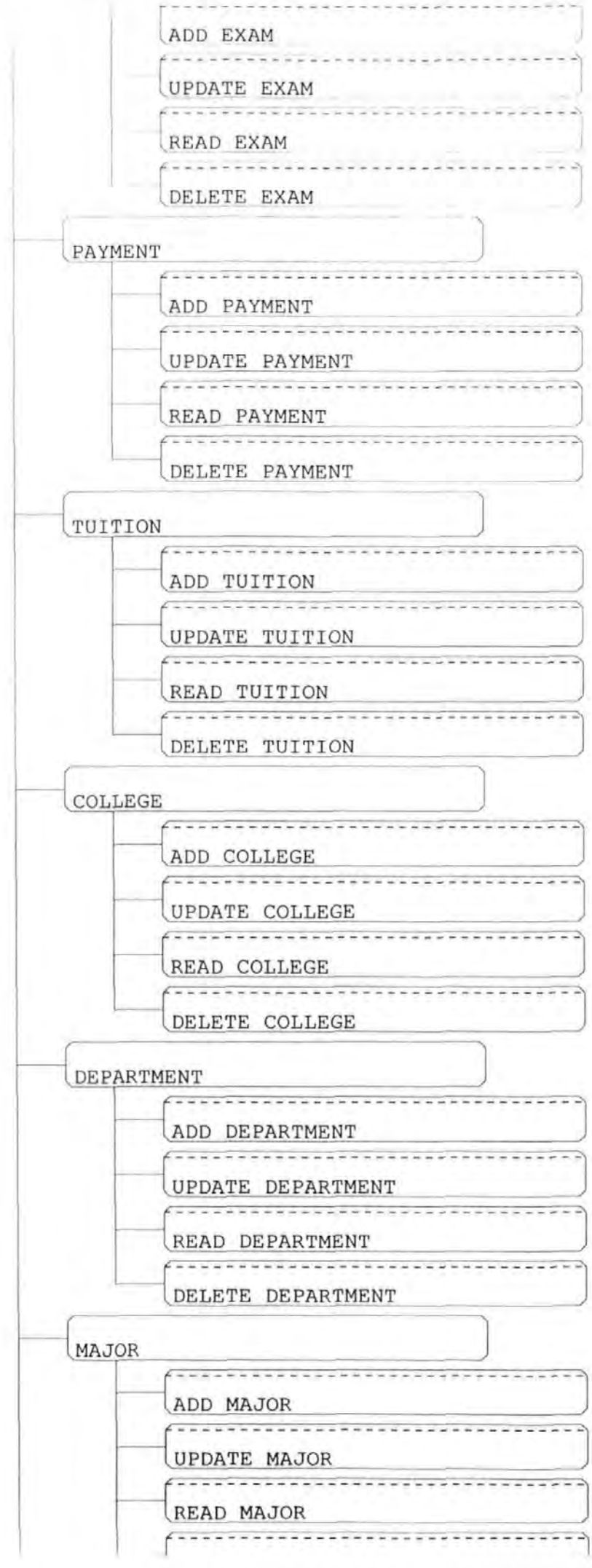
Cell Values:
= Not referenced
X = Include

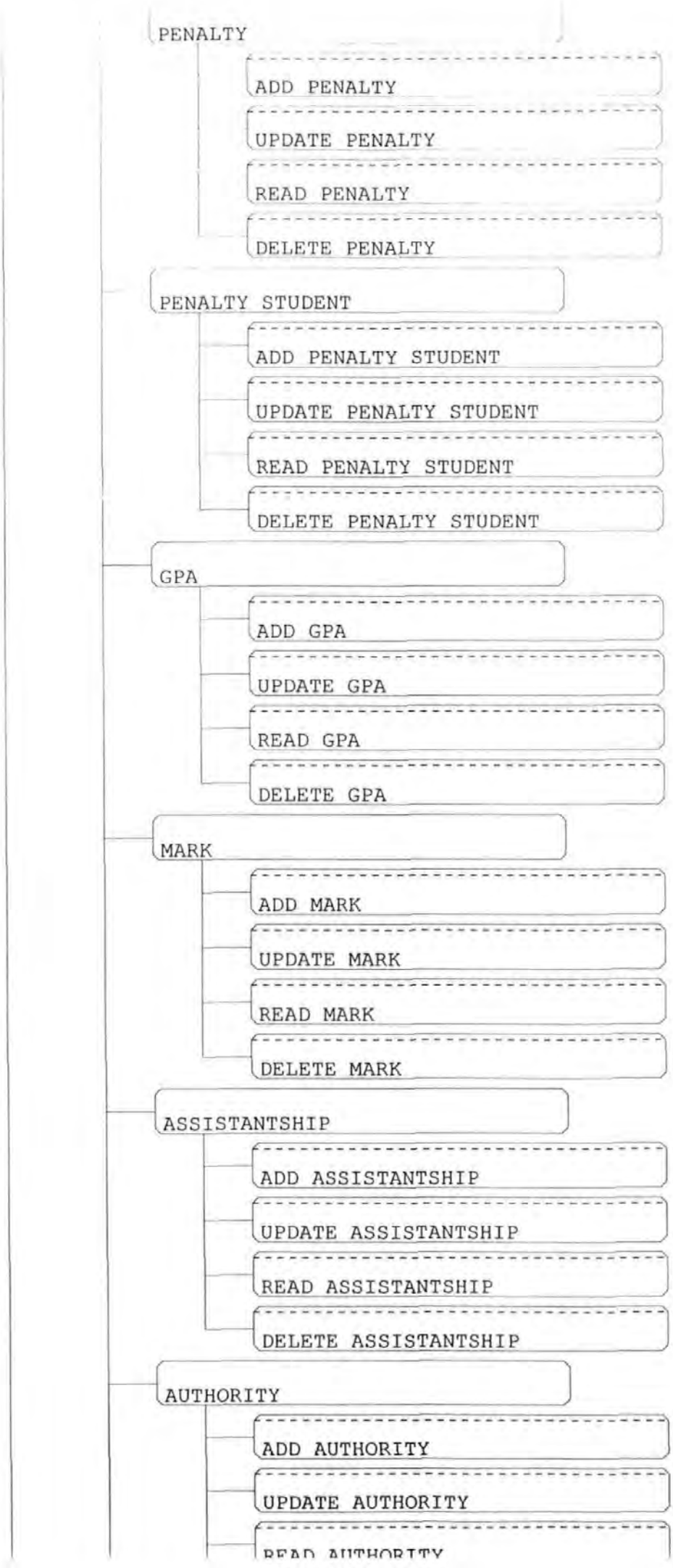
1 2
4 5
7 8

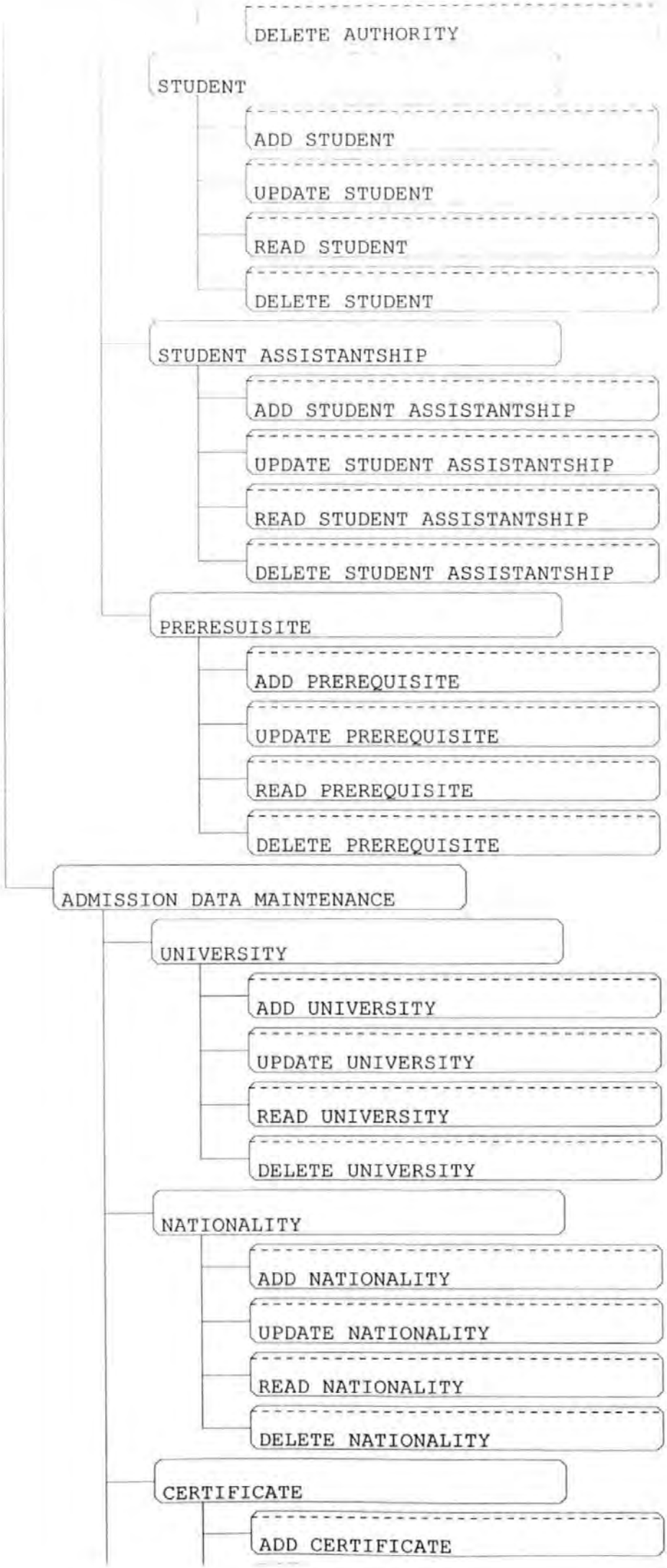
Elementary Process	Business System: ADMISSION REGISTRATION
DELETE DEPARTMENT	X
ADD MAJOR	X
UPDATE MAJOR	X
READ MAJOR	X
DELETE MAJOR	X
ADD NATIONALITY	X
UPDATE NATIONALITY	X
READ NATIONALITY	X
DELETE NATIONALITY	X
ADD PENALTY	X
UPDATE PENALTY	X
READ PENALTY	X
DELETE PENALTY	X
ADD PENALTY STUDENT	X
UPDATE PENALTY STUDENT	X
READ PENALTY STUDENT	X
DELETE PENALTY STUDENT	X
ADD MARK	X
UPDATE MARK	X
READ MARK	X
DELETE MARK	X
ADD ASSISTANTSHIP	X
UPDATE ASSISTANTSHIP	X
READ ASSISTANTSHIP	X
DELETE ASSISTANTSHIP	X
ADD AUTHORITY	X
UPDATE AUTHORITY	X
READ AUTHORITY	X
DELETE AUTHORITY	X
ADD STUDENT	X
UPDATE STUDENT	X
READ STUDENT	X
DELETE STUDENT	X
ADD STUDENT ASSISTANTSHIP	X
UPDATE STUDENT ASSISTANTSHIP	X
READ STUDENT ASSISTANTSHIP	X
DELETE STUDENT ASSISTANTSHIP	X
ADD GPA	X
UPDATE GPA	X
READ GPA	X
DELETE GPA	X
ADD UNIVERSITY	X
UPDATE UNIVERSITY	X
READ UNIVERSITY	X
DELETE UNIVERSITY	X
ADD PREREQUISITE	X
UPDATE PREREQUISITE	X
READ PREREQUISITE	X
DELETE PREREQUISITE	X

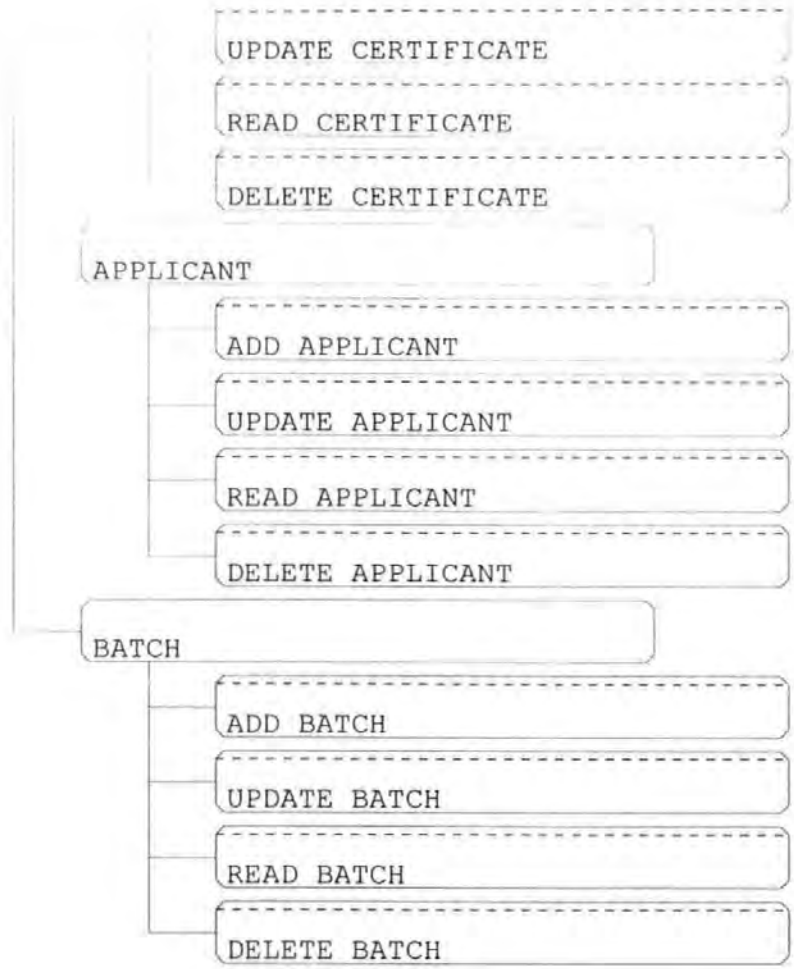
5- Activity Hierarchy



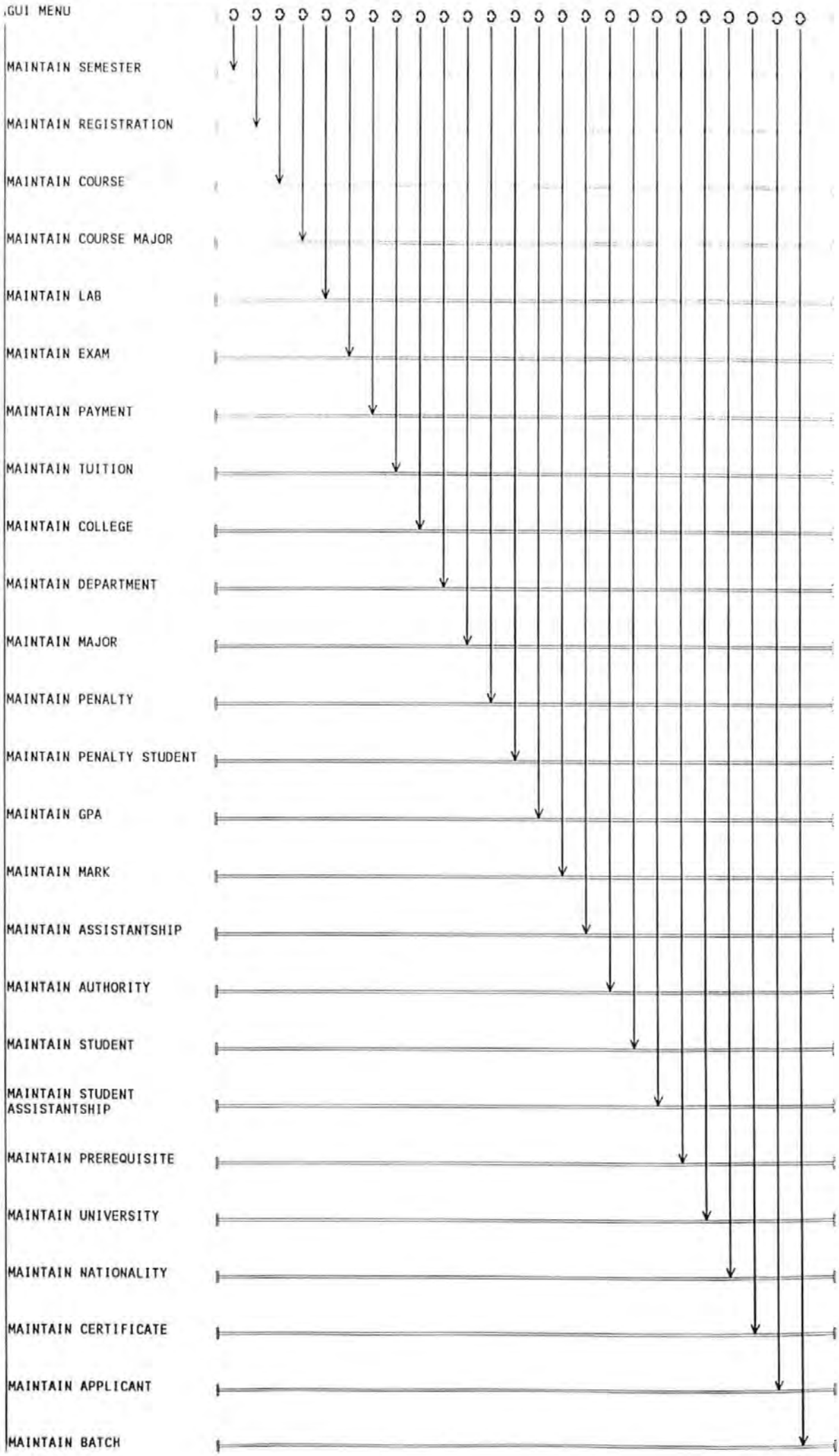

ADD EXAM
UPDATE EXAM
READ EXAM
DELETE EXAM
PAYMENT
ADD PAYMENT
UPDATE PAYMENT
READ PAYMENT
DELETE PAYMENT
TUITION
ADD TUITION
UPDATE TUITION
READ TUITION
DELETE TUITION
COLLEGE
ADD COLLEGE
UPDATE COLLEGE
READ COLLEGE
DELETE COLLEGE
DEPARTMENT
ADD DEPARTMENT
UPDATE DEPARTMENT
READ DEPARTMENT
DELETE DEPARTMENT
MAJOR
ADD MAJOR
UPDATE MAJOR
READ MAJOR


PENALTY
ADD PENALTY
UPDATE PENALTY
READ PENALTY
DELETE PENALTY
PENALTY STUDENT
ADD PENALTY STUDENT
UPDATE PENALTY STUDENT
READ PENALTY STUDENT
DELETE PENALTY STUDENT
GPA
ADD GPA
UPDATE GPA
READ GPA
DELETE GPA
MARK
ADD MARK
UPDATE MARK
READ MARK
DELETE MARK
ASSISTANTSHIP
ADD ASSISTANTSHIP
UPDATE ASSISTANTSHIP
READ ASSISTANTSHIP
DELETE ASSISTANTSHIP
AUTHORITY
ADD AUTHORITY
UPDATE AUTHORITY
READ AUTHORITY


DELETE AUTHORITY
STUDENT
ADD STUDENT
UPDATE STUDENT
READ STUDENT
DELETE STUDENT
STUDENT ASSISTANTSHIP
ADD STUDENT ASSISTANTSHIP
UPDATE STUDENT ASSISTANTSHIP
READ STUDENT ASSISTANTSHIP
DELETE STUDENT ASSISTANTSHIP
PRERESUISITE
ADD PREREQUISITE
UPDATE PREREQUISITE
READ PREREQUISITE
DELETE PREREQUISITE
ADMISSION DATA MAINTENANCE
UNIVERSITY
ADD UNIVERSITY
UPDATE UNIVERSITY
READ UNIVERSITY
DELETE UNIVERSITY
NATIONALITY
ADD NATIONALITY
UPDATE NATIONALITY
READ NATIONALITY
DELETE NATIONALITY
CERTIFICATE
ADD CERTIFICATE


UPDATE CERTIFICATE
READ CERTIFICATE
DELETE CERTIFICATE
APPLICANT
ADD APPLICANT
UPDATE APPLICANT
READ APPLICANT
DELETE APPLICANT
BATCH
ADD BATCH
UPDATE BATCH
READ BATCH
DELETE BATCH

6- Dialog Flow Diagram



7-Process/Procedure Action Diagrams (PAD/PrAD)[2]
-WINDOW APPLICANT

```
          Window/Dialog Definition

Business System   ADMISSION_REGISTRATION_DSS
Procedure         MAINTAIN_APPLICANT
Procedure Step    MAINTAIN_APPLICANT

Type:    Dialog Box

Title:   MAINTAIN_APPLICANT

Properties:
    Initial Position: Mouse Alignment
    Modality:         Application Modal
    Background:       Scaled
                      Bitmap:  <None>
    Style:            System Menu   Dialog Border
    Icon File:        <None>
    Name:             D1MAINTAIN_APPLICANT
    Dialect:          DEFAULT

Video Properties:
    Foreground Color:
        Red = 255   Green = 128   Blue = 128
    Background Color:
        Red = 128   Green = 255   Blue = 255
    Font Selection:
        Name:      Tahoma
        Style:     Bold
        Size:      10
        Emphasis:  <None>

Menu Design:         <Not applicable>

Status Bar:          <Not applicable>

Toolbar:             <Not applicable>

Window Controls:

    Hidden Field(s)
        <No Hidden Fields Defined>

    Push Button
        Text:                Help
        Mnemonic Key:        H
        Bitmap:              <None>
        Default Push Button not enabled
        Command:             <None>


        Button Action:       Special  Action...
            Help     - Execute Help system
        Disabling States:    <None>
        Video Properties:
            Foreground Color:
```

[2] Not all **PAD/PrAD** will be listed here, again for space reasons.

```
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 192   Green = 192   Blue = 192
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 3       Top   = 16
        Left   = 180     Right = 228

Push Button
    Text:                Ok
    Mnemonic Key:        O
    Bitmap:              <None>
    Default Push Button enabled
    Command:             <None>
    Button Action:       Special  Action...
        Ok        - Execute and Close
    Disabling States:    <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 192   Green = 192   Blue = 192
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 3       Top   = 16
        Left   = 3       Right = 39

Push Button
    Text:                Cancel
    Mnemonic Key:        C
    Bitmap:              <None>
    Default Push Button not enabled
    Command:             <None>
    Button Action:       Special  Action...
        Cancel - Close without Execution


    Disabling States:    <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 192   Green = 192   Blue = 192
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 3       Top   = 16
        Left   = 77      Right = 137

Entry Field
```

```
    Import View:     IMPORT
    Export View:     EXPORT
    Entity:          STUDENT
    Attribute:       STU_REGISTRATION_NUMBER
    Prompt:          Stu Registration Number:
    Edit Pattern:    <Default Edit Pattern>
    Edit Align:      Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 22      Top    = 29
        Left   = 159     Right = 219

Entry Field
    Import View:     IMPORT
    Export View:     EXPORT

    Entity:          CERTIFICATE
    Attribute:       CER_NUMBER
    Prompt:          Cer Number:
    Edit Pattern:    <Default Edit Pattern>
    Edit Align:      Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 34      Top    = 41
        Left   = 159     Right = 177

Entry Field
```

```
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         MAJOR
    Attribute:      MAJ_SNO
    Prompt:         Maj Sno:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255

        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 46     Top   = 53
        Left   = 159    Right = 177

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         BATCH
    Attribute:      BAT_NUMBER
    Prompt:         Bat Number:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 58     Top   = 65
        Left   = 159    Right = 183

Entry Field
```

```
    Import View:     IMPORT
    Export View:     EXPORT
    Entity:          UNIVERSITY
    Attribute:       UNI_ID
    Prompt:          Uni Id:
    Edit Pattern:    <Default Edit Pattern>
    Edit Align:      Default
    Properties:

        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 70      Top   = 77
        Left   = 159     Right = 177

Entry Field
    Import View:     IMPORT
    Export View:     EXPORT
    Entity:          NATIONALITY
    Attribute:       NAT_IDENTIFICATION
    Prompt:          Nat Identification:
    Edit Pattern:    <Default Edit Pattern>
    Edit Align:      Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 82      Top   = 89
        Left   = 159     Right = 177
```

```
    Group Box
        Group Box Text:     App Interviewed:
        Video Properties:
            Foreground Color:
                Red = 255   Green = 128   Blue = 128
            Background Color:
                Red = 128   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 93     Top   = 113
            Left   = 159    Right = 227
        Import View:    IMPORT
        Export View:    EXPORT
        Entity:         APPLICANT
        Attribute:      APP_INTERVIEWED
        Radio Buttons:
            Value:              f
            Prompt:             f
            Mnemonic Key:       f
            Disabling States:   <None>
            Video Properties:
                Foreground Color:
                    Red = 255   Green = 0   Blue = 0
                Background Color:
                    Red = 128   Green = 255   Blue = 255
                Font Selection:
                    Name:       Tahoma
                    Style:      Bold
                    Size:       10
                    Emphasis:   <None>
            Coordinates:
                Bottom = 94     Top   = 106
                Left   = 165    Right = 193
            -----

            Value:              p
            Prompt:             p
            Mnemonic Key:       p
            Disabling States:   <None>
            Video Properties:
                Foreground Color:
                    Red = 255   Green = 0   Blue = 0
                Background Color:
                    Red = 128   Green = 255   Blue = 255
                Font Selection:
                    Name:       Tahoma
                    Style:      Bold
                    Size:       10
                    Emphasis:   <None>
            Coordinates:
                Bottom = 94     Top   = 106
                Left   = 193    Right = 221

Defined Events:     <None>

Business System   ADMISSION_REGISTRATION_DSS
Procedure         MAINTAIN_APPLICANT
```

```
Procedure Step    MAINTAIN_APPLICANT

Type:    Primary Window

Title:   MAINTAIN_APPLICANT

Properties:
    Initial Position: Mouse Alignment
    Modality:         Application Modal
    Background:       Scaled
                      Bitmap:  <None>
    Style:            System Menu   Minimize Button   Maximize Button
    Icon File:        <None>
    Name:             MAINTAIN_APPLICANT
    Dialect:          DEFAULT

Video Properties:
    Foreground Color:
        Red = 128   Green = 0   Blue = 0
    Background Color:
        Red = 0   Green = 255   Blue = 255
    Font Selection:
        Name:      Tahoma
        Style:     Bold
        Size:      10
        Emphasis:  <None>

Menu Design:
    Main Menu Bar
                       Video Properties:
                       Foreground Color:
                           Red = 0   Green = 0   Blue = 0
                       Background Color:
                           Red = 192   Green = 192   Blue = 192
                       Font Selection:
                           Name:      Tahoma
                           Style:     Bold
                           Size:      10
                           Emphasis:  <None>

    --applicant
                      Sub Menu Item Properties:
                          Not Preceded by a separator line
                          Mnemonic Key:
                      Disabling States:   <None>


    ----add
                      Menu Item Properties:
                          Not Preceded by a separator line
                          Mnemonic Key:   _
                          Command:        CREATE
                          Accelerator :   F1
                          Button Action:  Executes the Procedure Step
                      Disabling States:   <None>
                      Defined Events:     <None>

    ----read
                      Menu Item Properties:
                          Not Preceded by a separator line
                          Mnemonic Key:   _
```

```
                        Command:        DISPLAY
                        Accelerator :   F2
                        Button Action:  Executes the Procedure Step
                    Disabling States:   <None>
                    Defined Events:     <None>

    ----update
                    Menu Item Properties:
                        Not Preceded by a separator line
                        Mnemonic Key:
                        Command:        UPDATE
                        Accelerator :   F3
                        Button Action:  Executes the Procedure Step
                    Disabling States:   <None>
                    Defined Events:     <None>

    ----delete
                    Menu Item Properties:
                        Not Preceded by a separator line
                        Mnemonic Key:
                        Command:        DELETE
                        Accelerator :   F4
                        Button Action:  Executes the Procedure Step
                    Disabling States:   <None>
                    Defined Events:     <None>

    --More...
                    Menu Item Properties:
                        Not Preceded by a separator line
                        Mnemonic Key:   M
                        Command:        <None>
                        Accelerator :   <None>
                        Button Action:  Initiates the Dialog Box...

                            D1MAINTAIN_APPLICANT
                    Disabling States:   <None>
                    Defined Events:     <None>

Status Bar:         <None>

Toolbar:            <None>

Window Controls:

    Hidden Field(s)
        <No Hidden Fields Defined>

    Entry Field
        Import View:    IMPORT
        Export View:    EXPORT
        Entity:         APPLICANT
        Attribute:      APP_ACCEPTED_A
        Prompt:         APP_ACCEPTED_A:
        Edit Pattern:   <Default Edit Pattern>
        Edit Align:     Default
        Properties:
            Autoscroll enabled
            Read Only not enabled
            Auto Tab not enabled
            Margin Box enabled
            Upper Case not enabled
```

```
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 47      Top   = 56
        Left   = 385     Right = 392

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_HOLD_TO_NEXT_BATCH_G

    Prompt:         APP_HOLD_TO_NEXT_BATCH_G:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 46      Top   = 55
        Left   = 270     Right = 278

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_PREDICTED_MAJOR_F
    Prompt:         APP_PREDICTED_MAJOR_F:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
```

```
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma

            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 11      Top   = 20
        Left   = 144     Right = 337

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_PREDICTED_GRADUATION_GRADE_M
    Prompt:         APP_PREDICTED_GRADUATION_GRADE_M:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 26      Top   = 35
        Left   = 205     Right = 335

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_AGE
    Prompt:         APP_AGE:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
```

```
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 46      Top   = 55
        Left   = 86      Right = 108

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_TELEPHONE
    Prompt:         App Telephone:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 63      Top   = 70
        Left   = 197     Right = 287

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_CERTIFICATION_PERCENT
    Prompt:         App Certification Percent:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
```

```
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:      Tahoma
            Style:     Bold
            Size:      10
            Emphasis:  <None>
    Coordinates:
        Bottom = 75      Top   = 82
        Left   = 197     Right = 222

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_CERTIFICATE_YEAR
    Prompt:         App Certificate Year:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:      Tahoma
            Style:     Bold
            Size:      10
            Emphasis:  <None>
    Coordinates:
        Bottom = 87      Top   = 94
        Left   = 197     Right = 222

Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_CODE
    Prompt:         App Code:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
```

```
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 99      Top   = 106
        Left   = 197     Right = 234

Group Box
    Group Box Text:      App Transferred:
    Video Properties:
        Foreground Color:
            Red = 128   Green = 0   Blue = 0
        Background Color:
            Red = 0   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 110     Top   = 130
        Left   = 197     Right = 265

    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_TRANSFERRED
    Radio Buttons:
        Value:              y
        Prompt:             y
        Mnemonic Key:       y
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 255   Green = 0   Blue = 0
            Background Color:
                Red = 128   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 111     Top   = 123
            Left   = 204     Right = 232
        -----

        Value:              n
        Prompt:             n
        Mnemonic Key:       n
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 255   Green = 0   Blue = 0
```

```
            Background Color:
                Red = 128   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 111    Top   = 123
            Left   = 232    Right = 259
        -----


Group Box
    Group Box Text:     App Previously Abandoned:
    Video Properties:
        Foreground Color:
            Red = 128   Green = 0   Blue = 0
        Background Color:
            Red = 0   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 135    Top   = 155
        Left   = 197    Right = 265

    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_PREVIOUSLY_ABANDONED
    Radio Buttons:
        Value:              y
        Prompt:             y
        Mnemonic Key:       y
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 255   Green = 0   Blue = 0
            Background Color:
                Red = 128   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 136    Top   = 148
            Left   = 204    Right = 232
        -----

        Value:              n
        Prompt:             n
        Mnemonic Key:       n
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 255   Green = 0   Blue = 0
            Background Color:
```

```
            Red = 128   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 136    Top    = 148
        Left   = 232    Right  = 259
    -----


Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_ADDRESS
    Prompt:         App Address:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
            Red = 255   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 161    Top    = 168
        Left   = 197    Right  = 407

Group Box
    Group Box Text:     App Gender:
    Video Properties:
        Foreground Color:
            Red = 128   Green = 0   Blue = 0
        Background Color:
            Red = 0   Green = 255   Blue = 255
        Font Selection:
            Name:       Tahoma
            Style:      Bold
            Size:       10
            Emphasis:   <None>
    Coordinates:
        Bottom = 172    Top    = 192
        Left   = 197    Right  = 265

    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
```

```
    Attribute:      APP_GENDER
    Radio Buttons:
        Value:              f
        Prompt:             f
        Mnemonic Key:       f
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 255   Green = 0   Blue = 0
            Background Color:
                Red = 128   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 173    Top   = 185
            Left   = 204    Right = 232
        -----

        Value:              m
        Prompt:             m
        Mnemonic Key:       m
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 255   Green = 0   Blue = 0
            Background Color:
                Red = 128   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 173    Top   = 185
            Left   = 232    Right = 259
        -----


Entry Field
    Import View:    IMPORT
    Export View:    EXPORT
    Entity:         APPLICANT
    Attribute:      APP_DOB
    Prompt:         App Dob:
    Edit Pattern:   <Default Edit Pattern>
    Edit Align:     Default
    Properties:
        Autoscroll enabled
        Read Only not enabled
        Auto Tab not enabled
        Margin Box enabled
        Upper Case not enabled
        Password not enabled
    Disabling States:   <None>
    Video Properties:
        Foreground Color:
            Red = 0   Green = 0   Blue = 0
        Background Color:
```

```
                Red = 255   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 198    Top   = 205
            Left   = 197    Right = 245

    Entry Field
        Import View:    IMPORT
        Export View:    EXPORT
        Entity:         APPLICANT
        Attribute:      APP_FULL_NAME

        Prompt:         App Full Name:
        Edit Pattern:   <Default Edit Pattern>
        Edit Align:     Default
        Properties:
            Autoscroll enabled
            Read Only not enabled
            Auto Tab not enabled
            Margin Box enabled
            Upper Case not enabled
            Password not enabled
        Disabling States:   <None>
        Video Properties:
            Foreground Color:
                Red = 0   Green = 0   Blue = 0
            Background Color:
                Red = 255   Green = 255   Blue = 255
            Font Selection:
                Name:       Tahoma
                Style:      Bold
                Size:       10
                Emphasis:   <None>
        Coordinates:
            Bottom = 210    Top   = 217
            Left   = 197    Right = 407
Defined Events:     <None>

 MAINTAIN_APPLICANT View Mapping

                  Export Views
     Supplying Import Views
    view of         EXPORT                                          to
     IMPORT
    entity           APPLICANT
     APPLICANT
    attr              APP_FULL_NAME
     APP_FULL_NAME
    attr              APP_DOB
     APP_DOB
    attr              APP_GENDER
     APP_GENDER
    attr              APP_ADDRESS
     APP_ADDRESS
    attr              APP_PREVIOUSLY_ABANDONED
     APP_PREVIOUSLY_ABANDONED
    attr              APP_TRANSFERRED
```

```
 APP_TRANSFERRED
attr              APP_CODE
 APP_CODE
attr              APP_CERTIFICATE_YEAR
 APP_CERTIFICATE_YEAR
attr              APP_CERTIFICATION_PERCENT
 APP_CERTIFICATION_PERCENT
attr              APP_TELEPHONE
 APP_TELEPHONE
attr              APP_INTERVIEWED
 APP_INTERVIEWED
attr              APP_AGE
 APP_AGE
attr              APP_PREDICTED_GRADUATION_GRADE_M
 APP_PREDICTED_GRADUATION_GRADE_M
attr              APP_PREDICTED_MAJOR_F
 APP_PREDICTED_MAJOR_F
attr              APP_HOLD_TO_NEXT_BATCH_G
 APP_HOLD_TO_NEXT_BATCH_G
attr              APP_ACCEPTED_A
 APP_ACCEPTED_A
view of         EXPORT                                          to
 IMPORT
entity           NATIONALITY
 NATIONALITY
attr              NAT_IDENTIFICATION
 NAT_IDENTIFICATION
view of         EXPORT                                          to
 IMPORT
entity           UNIVERSITY
 UNIVERSITY
attr              UNI_ID
 UNI_ID
view of         EXPORT                                          to
 IMPORT
entity           BATCH
 BATCH
attr              BAT_NUMBER
 BAT_NUMBER
view of         EXPORT                                          to
 IMPORT
entity           MAJOR
 MAJOR
attr              MAJ_SNO
 MAJ_SNO
view of         EXPORT                                          to
 IMPORT
entity           CERTIFICATE
 CERTIFICATE
attr              CER_NUMBER
 CER_NUMBER
view of         EXPORT                                          to
 IMPORT
entity           STUDENT
 STUDENT
attr              STU_REGISTRATION_NUMBER
 STU_REGISTRATION_NUMBER
view of         EXPORT_HIDDEN_ID
entity           APPLICANT
attr              APP_CODE
```

8- Applicant entity type logic[3]

8-1 ACCEPT APPLICANT LOGIC

BAA Action Block: DA_APP_ACCEPTED

```
Action Block Description:

 +- DA_APP_ACCEPTED
 |    IMPORTS: ...
 |    EXPORTS: ...
 |    LOCALS: ...
 |    ENTITY ACTIONS: ...
 |
 |  +- READ required applicant
 |  |        WHERE DESIRED required applicant app_code IS EQUAL TO input
applicant
 |  |              app_code
 |  +- WHEN successful
 |  |  SUMMARIZE required applicant
 |  |             required batch
 |  |        PLACING count(OCCURRENCES) INTO loc ief_supplied count
 |  |        WHERE DESIRED required batch maintains CURRENT required
applicant
 |  |  +- READ required major
 |  |  |        WHERE DESIRED required major has CURRENT required applicant
 |  |  +- WHEN successful
 |  |  |  +- CASE OF required major maj_title
 |  |  |  +- CASE "b of hotels and tourism"
 |  |  |  |  +- IF required applicant app_interviewed IS EQUAL TO "p"
 |  |  |  |  |        AND required applicant app_previously_abandoned IS
EQUAL TO "n"
 |  |  |  |  |        AND required applicant app_certification_percent
 |  |  |  |  |              IS GREATER OR EQUAL TO 60
 |  |  |  |  |        AND year(CURRENT_DATE) - required applicant
app_certificate_year
 |  |  |  |  |              IS LESS OR EQUAL TO 2
 |  |  |  |  |        AND year(CURRENT_DATE) - year(required applicant
app_dob)
 |  |  |  |  |              IS LESS OR EQUAL TO 25
 |  |  |  |  |        AND required batch bat_closing_date IS GREATER THAN
CURRENT_DATE
 |  |  |  |  |        AND loc ief_supplied count IS LESS THAN required
batch bat_ceiling
 |  |  |  |  |  SET output applicant app_accepted_a TO "y"
 |  <------------ESCAPE
 |  |  |  |  +- ELSE
 |  |  |  |  |  SET output applicant app_accepted_a TO "n"
 |  |  |  |  +--
 |  |  |  +- CASE "bachelor of maritime"
 |  |  |  |  +- IF required applicant app_gender IS EQUAL TO "m"
 |  |  |  |  |        AND required applicant app_previously_abandoned IS
EQUAL TO "n"
 |  |  |  |  |        AND required applicant app_certification_percent
 |  |  |  |  |              IS GREATER OR EQUAL TO 60
 |  |  |  |  |        AND year(CURRENT_DATE) - required applicant
app_certificate_year
 |  |  |  |  |              IS LESS OR EQUAL TO 2
 |  |  |  |  |        AND year(CURRENT_DATE) - year(required applicant
app_dob)
```

[3] No all entity types' logic will be presented here because of the space constraint.

```
|  |  |  |  |  |              IS LESS OR EQUAL TO 25
|  |  |  |  |  |        AND required batch bat_closing_date IS GREATER THAN
CURRENT_DATE
|  |  |  |  |  |         AND loc ief_supplied count IS LESS THAN required
batch bat_ceiling
|  |  |  |  |  |  SET output applicant app_accepted_a TO "y"
|  |  |  |  |  |  MOVE required applicant  TO output applicant
|  <------------ESCAPE
|  |  |  |  +- ELSE
|  |  |  |  |  SET output applicant app_accepted_a TO "n"
|  |  |  |  +--
|  |  |  +- CASE "bachelor of marine eng"
|  |  |  |  +- IF required applicant app_gender IS EQUAL TO "m"
|  |  |  |  |  |        AND required applicant app_previously_abandoned IS
EQUAL TO "n"
|  |  |  |  |  |        AND required applicant app_certification_percent
|  |  |  |  |  |              IS GREATER OR EQUAL TO 60
|  |  |  |  |  |        AND year(CURRENT_DATE) - required applicant
app_certificate_year
|  |  |  |  |  |              IS LESS OR EQUAL TO 2
|  |  |  |  |  |        AND year(CURRENT_DATE) - year(required applicant
app_dob)
|  |  |  |  |  |              IS LESS OR EQUAL TO 25
|  |  |  |  |  |        AND required batch bat_closing_date IS GREATER THAN
CURRENT_DATE
|  |  |  |  |  |        AND loc ief_supplied count IS LESS THAN required
batch bat_ceiling
|  |  |  |  |  |  SET output applicant app_accepted_a TO "y"
|  <------------ESCAPE
|  |  |  |  +- ELSE
|  |  |  |  |  SET output applicant app_accepted_a TO "n"
|  |  |  |  +--
|  |  |  +- CASE "b tech marine eng"
|  |  |  |  +- IF required applicant app_gender IS EQUAL TO "m"
|  |  |  |  |  |        AND required applicant app_previously_abandoned IS
EQUAL TO "n"
|  |  |  |  |  |        AND required applicant app_certification_percent
|  |  |  |  |  |              IS GREATER OR EQUAL TO 60
|  |  |  |  |  |        AND year(CURRENT_DATE) - required applicant
app_certificate_year
|  |  |  |  |  |              IS LESS OR EQUAL TO 2
|  |  |  |  |  |        AND year(CURRENT_DATE) - year(required applicant
app_dob)
|  |  |  |  |  |              IS LESS OR EQUAL TO 25
|  |  |  |  |  |        AND required batch bat_closing_date IS GREATER THAN
CURRENT_DATE
|  |  |  |  |  |        AND loc ief_supplied count IS LESS THAN required
batch bat_ceiling
|  |  |  |  |  |  SET output applicant app_accepted_a TO "y"
|  <------------ESCAPE
|  |  |  |  +- ELSE
|  |  |  |  |  SET output applicant app_accepted_a TO "n"
|  |  |  |  +--
|  |  |  +- OTHERWISE
|  |  |  |  +- IF required applicant app_gender IS NOT EQUAL TO "x"
|  |  |  |  |  |        AND required applicant app_previously_abandoned IS
EQUAL TO "n"
|  |  |  |  |  |        AND required applicant app_certification_percent
|  |  |  |  |  |              IS GREATER OR EQUAL TO 60
|  |  |  |  |  |        AND year(CURRENT_DATE) - required applicant
app_certificate_year
```

```
 |  |  |  |  |              IS LESS OR EQUAL TO 2
 |  |  |  |  |         AND year(CURRENT_DATE) - year(required applicant
app_dob)
 |  |  |  |  |              IS LESS OR EQUAL TO 25
 |  |  |  |  |         AND required batch bat_closing_date IS GREATER THAN
CURRENT_DATE
 |  |  |  |  |         AND loc ief_supplied count IS LESS THAN required
batch bat_ceiling
 |  |  |  |  |  SET output applicant app_accepted_a TO "y"
 | <------------ESCAPE
 |  |  |  |  +- ELSE
 |  |  |  |  |  SET output applicant app_accepted_a TO "n"
 |  |  |  |  +--
 |  |  |  +--
 |  |  +- WHEN not found
 |  |  |  EXIT STATE IS no_matching_records
 |  |  +--
 |  +- WHEN not found
 |  |  EXIT STATE IS applicant_nf
 |  +--
 +--
```

8-2 ADD APPLICANT LOGIC

```
                                Process:  ADD_APPLICANT
Process Description:
Action Block Description:

 +- ADD_APPLICANT
 |    IMPORTS: ...
 |    EXPORTS: ...
 |    LOCALS:
 |    ENTITY ACTIONS: ...
 |
 |  +- READ certificate
 |  |       WHERE DESIRED certificate cer_number IS EQUAL TO import
certificate
 |  |             cer_number
 |  +- WHEN successful
 |  |  MOVE certificate  TO export certificate
 |  |  +- READ major
 |  |  |       WHERE DESIRED major maj_sno IS EQUAL TO import major
maj_sno
 |  |  +- WHEN successful
 |  |  |  MOVE major  TO export major
 |  |  |  +- READ batch
 |  |  |  |       WHERE DESIRED batch bat_number IS EQUAL TO import batch
bat_number
 |  |  |  +- WHEN successful
 |  |  |  |  MOVE batch  TO export batch
 |  |  |  |  +- CREATE applicant
 |  |  |  |  |  ASSOCIATE  WITH batch WHICH maintains IT
 |  |  |  |  |  ASSOCIATE  WITH major WHICH has IT
 |  |  |  |  |  ASSOCIATE  WITH certificate WHICH points_to IT
 |  |  |  |  |  SET app_full_name TO import applicant app_full_name
 |  |  |  |  |  SET app_dob TO import applicant app_dob
 |  |  |  |  |  SET app_gender TO import applicant app_gender
 |  |  |  |  |  SET app_address TO import applicant app_address
 |  |  |  |  |  SET app_previously_abandoned TO import applicant
 |  |  |  |  |                app_previously_abandoned
 |  |  |  |  |  SET app_transferred TO import applicant app_transferred
 |  |  |  |  |  SET app_code TO import applicant app_code
 |  |  |  |  |  SET app_certificate_year TO import applicant
app_certificate_year
 |  |  |  |  |  SET app_certification_percent TO import applicant
 |  |  |  |  |                app_certification_percent
 |  |  |  |  |  SET app_telephone TO import applicant app_telephone
 |  |  |  |  |  SET app_interviewed TO import applicant app_interviewed
 |  |  |  |  +- WHEN successful
 |  |  |  |  |  MOVE applicant  TO export applicant
 |  |  |  |  |  +- READ university
 |  |  |  |  |  |       WHERE DESIRED university uni_id IS EQUAL TO import
university
 |  |  |  |  |  |             uni_id
 |  |  |  |  |  +- WHEN successful
 |  |  |  |  |  |  ASSOCIATE university
 |  |  |  |  |  |         WITH applicant WHICH transferred_from IT
 |  |  |  |  |  |  MOVE university  TO export university
 |  |  |  |  |  +- WHEN not found
 |  |  |  |  |  |  EXIT STATE IS university_nf
 |  |  |  |  |  +--
 |  |  |  |  |  +- READ nationality
```

```
 |  |  |  |  |  |        WHERE DESIRED nationality nat_identification IS
EQUAL TO
 |  |  |  |  |  |              import nationality nat_identification
 |  |  |  |  |  +- WHEN successful
 |  |  |  |  |  |  ASSOCIATE nationality
 |  |  |  |  |  |           WITH applicant WHICH has IT
 |  |  |  |  |  |  MOVE nationality  TO export nationality
 |  |  |  |  |  +- WHEN not found
 |  |  |  |  |  |  EXIT STATE IS nationality_nf
 |  |  |  |  |  +--
 |  |  |  |  +- WHEN already exists
 |  |  |  |  |  EXIT STATE IS applicant_ae
 |  |  |  |  +- WHEN permitted value violation
 |  |  |  |  |  EXIT STATE IS applicant_pv
 |  |  |  |  +--
 |  |  |  +- WHEN not found
 |  |  |  |  EXIT STATE IS batch_nf
 |  |  |  +--
 |  |  +- WHEN not found
 |  |  |  EXIT STATE IS major_nf
 |  |  +--
 |  +- WHEN not found
 |  |  EXIT STATE IS certificate_nf
 |  +--
 +--
```

8-3 DELETE APPLICANT LOGIC

Process: DELETE_APPLICANT

```
Process Description:
Action Block Description:

 +- DELETE_APPLICANT
 |    IMPORTS: ...
 |    EXPORTS: ...
 |    LOCALS:
 |    ENTITY ACTIONS: ...
 |
 |  +- READ applicant
 |  |        WHERE DESIRED applicant app_code IS EQUAL TO import applicant
app_code
 |  +- WHEN successful
 |  |  MOVE applicant  TO export applicant
 |  |  += READ EACH student
 |  |  |        WHERE DESIRED student was CURRENT applicant
 |  |  |  MOVE student  TO export student
 |  |  |  EXIT STATE IS applicant_cannot_be_dele
 |  | <---ESCAPE
 |  |  +--
 |  |  +- IF EXITSTATE IS EQUAL TO applicant_cannot_be_dele
 |  |  |  NOTE Payment cannot be deleted until no more student associated
<---------ESCAPE
 |  |  +- ELSE
 |  |  |  DELETE applicant
 |  |  +--
 |  +- WHEN not found
 |  |  EXIT STATE IS applicant_cannot_be_dele
 |  +--
 +--
```

8-4 READ APPLICANT LOGIC

```
                              Process:  READ_APPLICANT
Process Description:
Action Block Description:

 +- READ_APPLICANT
 |    IMPORTS: ...
 |    EXPORTS: ...
 |    LOCALS:
 |    ENTITY ACTIONS: ...
 |
 |  +- READ applicant
 |  |       WHERE DESIRED applicant app_code IS EQUAL TO import applicant
app_code
 |  +- WHEN successful
 |  |  MOVE applicant  TO export applicant
 |  +- WHEN not found
 |  |  EXIT STATE IS applicant_nf
 |  +--
 +--
```

8-5 - UPDATE APPLICANT LOGIC

```
                              Process:  UPDATE_APPLICANT
Process Description:
Action Block Description:

 +- UPDATE_APPLICANT
 |    IMPORTS: ...
 |    EXPORTS: ...
 |    LOCALS:
 |    ENTITY ACTIONS: ...
 |
 |  +- READ applicant
 |  |        WHERE DESIRED applicant app_code IS EQUAL TO import applicant
app_code
 |  +- WHEN successful
 |  |  +- UPDATE applicant
 |  |  |  SET app_full_name TO import applicant app_full_name
 |  |  |  SET app_dob TO import applicant app_dob
 |  |  |  SET app_gender TO import applicant app_gender
 |  |  |  SET app_address TO import applicant app_address
 |  |  |  SET app_previously_abandoned TO import applicant
app_previously_abandoned
 |  |  |  SET app_transferred TO import applicant app_transferred
 |  |  |  SET app_certificate_year TO import applicant app_certificate_year
 |  |  |  SET app_certification_percent TO import applicant
app_certification_percent
 |  |  |  SET app_telephone TO import applicant app_telephone
 |  |  |  SET app_interviewed TO import applicant app_interviewed
 |  |  +- WHEN successful
 |  |  |  MOVE applicant  TO export applicant
 |  |  +- WHEN not unique
 |  |  |  EXIT STATE IS applicant_nu
 |  |  +- WHEN permitted value violation
 |  |  |  EXIT STATE IS applicant_pv
 |  |  +--
 |  +- WHEN not found
 |  |  EXIT STATE IS applicant_nf
 |  +--
 +--
```

Appendix (G)

Conference papers

Knowledge Discovery in Database Techniques Applied to Students Recruitment Systems in Universities[1]

Ahmed El-Ragal
Arab Academy for Science and Technology
College of Management and Technology
P.O. Box 1029, AAST, Miami,
Alexandria, Egypt
Tel : +2-03-5485473
Fax : +2-03-5566072
E-mail: a.el-ragal@computer.org

Terry Mangles
University of Plymouth
Business School
Drake Circus, Plymouth, PL4 8AA
Devon, UK
Tel : +44-01752-232856
Fax : +44-01752-232853
E-mail : terry.mangles@pbs.plym.ac.uk

- **Abstract**

This paper will introduce the knowledge discovery in database (KDD) process applied to the students recruitment systems in universities. The definition of the KDD process and its importance will be defined. Different terminology for the knowledge discovery process will be discussed with particular emphasis on data mining. The distinction between KDD and data mining will be clarified by showing the place of the data mining in the KDD process. The tasks, goals, and components of the data mining algorithms are illustrated. The data mining techniques including query tools, visualization, on line analytical processing (OLAP), association rules, decision trees and rules, artificial neural networks (ANN), clustering, genetic algorithms and probabilistic graphical dependency technique will be discussed. The role of the user in the KDD process will be discussed. To place the entire KDD process in context, it is applied to a sample data set of 1600 records drawn from the Arab Academy for Science & Technology and Maritime Transport (AASTMT) records.

-**Key words**: knowledge discovery of database (KDD), data mining techniques, and student recruitment systems.

1- The emergence and definition of the KDD process

The term KDD was coined in 1989 to point to the process of finding knowledge in data (Fayyad, et al., 1996). KDD is also defined as the process of finding patterns hidden information or unknown facts in the database. Traditionally the notion of finding useful unknown patterns and hidden information in raw data has been given many titles including knowledge discovery in database, data mining, data archaeology, information discovery, knowledge discovery or extraction, and information harvesting (Adriaans and Zantinge, 1996). The reason for this lack of consensus has two reasons; the first is the novelty of the KDD and the second is the multi-disciplinary features of KDD. Multi-disciplinary means that KDD belongs to many disciplines like statistics and computer (machine learning, artificial intelligence (AI), databases, data warehousing, expert systems, knowledge acquisition and data visualization), from which the KDD process was drawn (Fayyad, et al., 1996). It is this broad applicability that has led number of researchers and scholars to have common interest in KDD.

The interest in KDD has been increased and this is demonstrated by the increasing number of forums and workshops. Another sources of interest are the various publications and special issues that document some of the KDD features and foundations (Inmon & Osterfelt 1991; Piatetsky-Sharipo 1992; Parsaye & Chignell 1993; Cercone & Tsuchiya 1993; Piatetsky-

[1] The 15th IAIM 2000 Conference, in Brisbane, Australia, 8-10 December 2000, pp. 7-20.

Sharipo 1995). With the Web's emergence as large distributed data store or repository and the realization that this large online data can de used for extensive commercial purposes, interest in the KDD process has widened. Although, the field of KDD was founded on many disciplines, however, it is gaining its character on its own and now stands by itself (Ramakrishnan and Grama, 1999).

2- KDD or data mining

Scientists have used the two terms KDD and data mining interchangeably (Ganti, et al., 1999). However, others said that data mining is a step in the KDD process (Adriaans and Zantinge, 1996). For the purpose of this paper data mining is considered a step in the KDD process. In other words, the research deals with KDD as the overall process of discovering useful knowledge from data while data mining points to the application algorithm or technique used for extracting patters and unknown information from the raw data. So the KDD process will get knowledge or information from the data mining techniques applied to a certain application.

3- The KDD process

The KDD process is interactive, iterative, and involves a great deal of user-interference. Brachman & Anand in 1996, defined the practical view of the KDD process as follows:

1. *Developing an understanding of the application domain.* This is an important step because it determines and the goals of the KDD application. Based on these goals the relevant data mining techniques can be employed, where there is no one single technique best fits all sources of application domains. In addition the over all performance of the KDD process will be evaluated based on the domain and its goal(s);
2. *Creating a target data set.* Selecting the data set, or focusing on a subset of the database on which discovery will take place;
3. *Data cleaning and preprocessing.* Basic operations such as the removal of noise if relevant, and deciding on the strategy of how to deal with missing data items. Example of strategies that might be used here are neglecting the incomplete data records or setting missing values to null. Since both affect the accuracy of the output knowledge this decision is important and the strategy used is often based on the volume of the incomplete data and its importance;
4. *Data reduction and projection.* Finding useful features to represent the data set depending on the goal of the task. For example if the goal of the KDD task is to determine and predict the students' academic performance, not all of the student's record is of importance. Examples are the students' address, telephone number or height and weight;
5. *Choosing the data mining task.* This is an important step in which the goal of the KDD process is defined as weather classification, clustering, summarization or others;
6. *Choosing the data mining technique(s) or algorithm(s).* Selecting the methods to be used for searching in the data for patterns and hidden information. This includes deciding on the relevant models and parameters. Also, there should be a match between the goal of the KDD and the data mining techniques or algorithms;
7. *Data miming.* Searching for patterns in the data sets using analysis methods and models such as regression, clustering, SQL, visualization, decision trees and others;
8. *Interpreting the information gained by the mining techniques.* The output of the mining techniques should be evaluated to be understandable and consistent. Iterations from steps 1 to 7 may also happen when apply KDD to real problem;
9. *Consolidating the discovered knowledge.* Reporting the knowledge to the interested parties and checking the discovered knowledge with the previously known knowledge.

During the KDD process, particularly the data mining step, it is necessary to search the database of the organization. Either the enterprise data warehouse (DW), or the data mart of

any department could enhance the search. So, data warehouse will enhance the KDD process through the wealth of historical information it offers to the mining techniques (Berson and Smith, 1997; Adriaans and Zanting, 1996). The following figure (1) describes the whole KDD process.

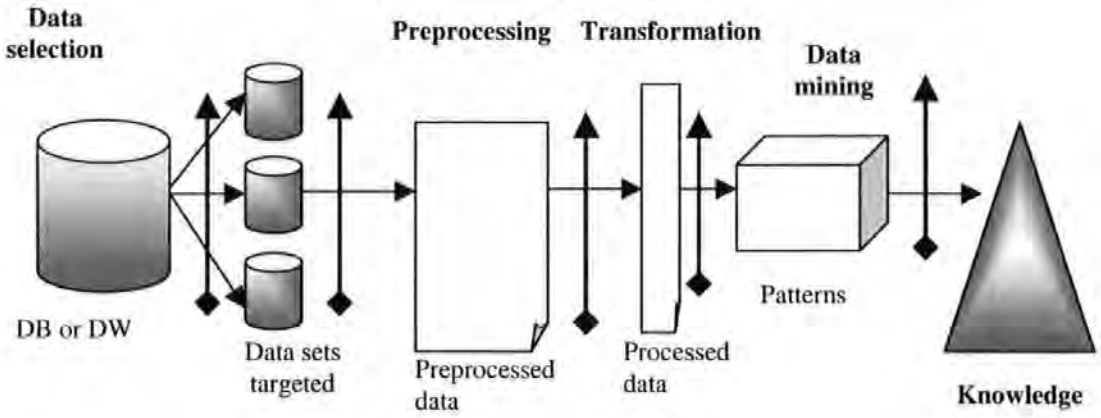


Figure (1). The KDD process overview[2].

The diagram indicates that the KDD is an entire process that should be applied from the application domain identification step until the evaluation of the discovered knowledge, after that the discovered knowledge should be utilized in a suitable front-end tool like executive information systems (EIS) or decision support systems (DSS). The data warehouse is a collection of data copied from other systems and assembled in one place, once it is assembled it became available to end-users who can use it to support different kinds of business decision support systems and information activities. When the DW contains all organisational data it is called enterprise DW, whilst if the DW contains functional data about marketing, personnel, or production it is said to be a data mart. The data warehouse (DW) will enhance the KDD results (Taha, et al., 1997; Berson and Smith, 1997; Barquin, 1997; Paller, 1997).

4- The primary tasks of data mining

The two high-level fundamental goals of data mining in practice tend to be *prediction* and *description* (Fayyad, et al., 1996). Prediction is the use of some variables or data fields in the database to predict the unknown future values of the other variables or data fields of interest. Description focuses on finding human understandable patterns describing the data set. The relative necessity of both of these goals varies from application to another, for some applications prediction is more important than description and for another applications description has higher importance. However, in the KDD process, description is much more important than prediction (Fayyad, et al., 1996). The case is always the opposite in the machine learning and pattern recognition applications, where prediction is the principal goal. Both goals, Prediction and description are achieved using the following data mining tasks.

1. *Classification*. It is a learning function that classifies a data into one of several predefined classes. Examples of classification methods used as part of knowledge discovery applications like classifying the students concerning the scholarship the university grants them into: granted and not-granted based on their GPA (grade point average);
2. *Regression*. A regression model is a mathematical equation that provides predictions of the values of one variable (dependent) based on the known values of one or more other variables (independent or predictors). If one predictor is considered then the regression is called simple linear regression, if more than one variable is used then the regression is

[2] Adapted from (Fayyad, et al., 1996).

called multiple linear regression (Canavos and Miller, 1995). Applications of regression are a lot, like predicting the new students that will join the university, predicting the census of a certain country, and loan predictions of a certain bank;

3. *Clustering.* Clustering, Q-analysis, typology, grouping, clumping, classification, numerical taxonomy and unsupervised pattern recognition are used interchangeably to refer to the same thing. Clustering is the process of producing classifications from initially unclassified data (Everitt, 1981). It is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. The categories may be mutually exclusive or exhaustive or may be overlapping (Fayyad et al., 1996). When the clusters overlap, this allows the data points to belong to one or more clusters;
4. *Summarization.* It is the process of finding a compact description for a subset of data. Examples of summarization like finding the mean and standard deviation of all data fields of interest, or the discovery of the relationships between variables, as well as multivariate visualization techniques. Summarization is applied to interactive exploratory data analysis and report generation;
5. *Dependency modelling.* It consists of finding a model that describes significant dependencies between variables. Dependency models fall into two categories: the structural category and the quantitative category. The structural one determines which variables are dependent on each other, while the quantitative one shows the strength of the dependencies using numerical scale. Example is *the probabilistic dependency networks* which are used in medical expert systems, modelling of the human genome, and information retrieval;
6. *Change and deviation detection.* It focuses on discovering the most significant changes in the data from previously measured or normative values.

5- Discussion of common data mining techniques

Any form that would help extracting more patterns and hidden information from the database is called a data mining technique (Adriaans and Zantinge, 1996). This is besides that data mining is a multi-disciplinary field in itself (Fayyad et al., 1996). These all means that there are more than one data mining technique, because again they have different backgrounds. Many classifications are found for data mining techniques based on many variant factors of classification (Ramakrishnan and Grama, 1999), like:

1. *The induced representation* (decision trees, rules, correlations, deviations, trends, and associations);
2. *The data they operate on* (time series, discrete, labelled, continuous, or nominal);
3. *The application domains* (finance, economic, biology, Web log mining and the like).

Another said that the data mining techniques are many (Adriaans and Zantinge, 1996), these are; Query tools, Visualization, On-line analytical processing (OLAP), Association rules, Case-based learning (nearest neighbour), Decision trees, Statistical techniques, Genetic algorithms (GA), Artificial neural networks (ANN), and Classification & regression techniques (CART). Discussions of the techniques that are of interest to this research are developed in the following sections.

5-1 Query tools

Traditional query tools are first used to analyze the data sets. By applying simple structured query language (SQL) we can obtain a wealth of information. However, before we can apply more advanced pattern analysis algorithms, we need to know some basic aspects and structures of the data set. With SQL we can uncover only shallow knowledge that is easily accessible from the data set: yet although we cannot find hidden knowledge. Adriaans and Zantinge (1996) said, "*for the most part 80% of the interesting information can be abstracted*

from a database using SQL. The remaining 20% of hidden requires more advanced techniques." However, for most organizations this 20% of the hidden knowledge have 80% of the importance in relation to decision making, and the 80% information volume represent only 20% in terms of value to the decision making process. A good way to start is to extract some simple statistical information from the data using SQL queries. For example:
-How many students in the university taking the accounting major?
-What is the average GPA for the male students with an American diploma background?
-How many grants go to junior students?
-What is the nationality distribution of students?
-What is the background of these students?
Decisions could be taken based on the output of the SQL statements.

5-2 Visualization

Visualization technique depends strongly on the human side of the analysis (Berson, 1996). Even the best set of rules or tables of data may reveal more information when visualized with color, relief, or texture in *2D, 3D* and even *4D* representations. In the 4D, 3D are mapped onto the screen and the fourth can be expressed through the use of color (Berson and Smith, 1997). Visualization technique may be used throughout the data exploration process and are particularly useful during the initial stages of the high-level groupings of data sets. For most users these advanced features are not accessible, and they have to rely on simple graphical display techniques that are contained in the query tool or data mining tools they are using. However, even these simple methods can provide us with a wealth of information. An elementary technique that can be of great value is the *scatter diagram*; in this technique, information on two attributes is displayed in a Cartesian space. Scatter diagrams can be used to identify interesting sub-sets of the data sets so that we can focus on the rest of the data mining process (Adriaans & Zantinge,1996).
Data visualization is emerging as an advanced technology that may allow organizations to process amounts of information and present it in a usable format. It is an interactive data manipulation technique on how to process a very huge amount of data. Colors, size, orientation, shape, and behavior can present multiple dimensions. Through the interface visualization techniques provide the non-computer users to navigate through data using their human feelings of the data displayed. Visualization technique plays a great role, that is it puts the information we have in an easy and understandable way for both computer and non-computer aware people (Keim, et al., 1996). As a result of this, managers can easily get the flavor of data through a deep looking to the diagrams provided by the technique.
The following benefits are earned beyond the use of visualization techniques (Berson, 1996):
-Users can easily interact with attributes and illustrate how they affect certain phenomenon;
-Users can view summarized data with drill down capability;
- Find hidden patterns.
-It is considered an exploratory data analysis tool (EDA). That is data navigation, comparison, scaling, filtering are all available to users;

5-3 On Line Analytical Processing (OLAP) tools

The need for OLAP was developed to handle situations where the relational database management systems (RDBMS) can not deal with *the multidimensional* problems. Although RDBMS are powerful solutions for a wide range of commercial and scientific applications, they are not good at addressing the modern business analysis, forecasting, and all the like that are multidimensional in nature (Berson, 1996). The key driver for OLAP is the multi-dimensional nature of the issues it deals with. For example managers might ask question like the following: Describe the relationship between majors, nationalities, ages, and GPA?

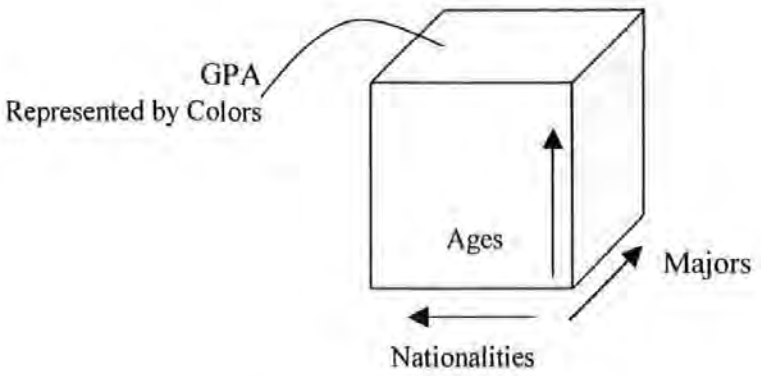


Figure (2). OLAP example.

Question like the previous one represents a 4-dimensional format. Although all the required information can be formulated using relational database and accessed via SQL, the two dimensional relational model of data and the SQL have some serious limitations for complex real world problems. Also response time and SQL functionality are a source of problems (Berson, 1996). OLAP is a continuos and iterative process, an analyst may drill down to see much more details. And this discovery of knowledge forces the managers to ask more complex questions.

Multidimensionality is the core of a number of OLAP systems available. However, there is a list of elements that determine the need for what product of OLAP to buy. Users should also prioritize this suggested list to reflect their business needs:

1. *Multidimensional conceptual view*. The tool must support the users with the level of dimensionality needed to reach some certain analysis;
2. *Transparency*. The heterogeneity of input data sources should be transparent to users to prevent their productivity from being decreased;
3. *Accessibility*. The OLAP system should access only the data required for analysis;
4. *Consistent reporting performance*. As the number of dimension increase and the database size also enlarged, users must expect the same performance;
5. *Client/Server architecture*. The OLAP system has to be compatible with the client/server architectural principles;
6. *Generic dimensionality*. Every data dimension should be in both its structure and operational capabilities;
7. *Multi-user support*. The OLAP system must be able to support multi-users working together;
8. *Flexible reporting*. The ability to arrange rows, columns, and cells in a way that facilitates visual analysis.

Several researchers have stated that OLAP is an independent technique and is as powerful as the data mining process and techniques (Fayyad, et al., 1996; Adriaans and Zantinge,1996). There is, however, an important difference between data mining and OLAP. OLAP is one of the data mining technique applied in the context of the data mining process. OLAP does not learn, unlike other data mining techniques, they also can not search for new solutions. OLAP also needs special storage media. Finally, data mining process, which entails many techniques, is more powerful than OLAP (Adriaans and Zantinge,1996; Berson,1996).

OLAP involves several basic analytical operations including consolidation, drill-down, and statistical techniques (O'Brien, 1996).

1. Consolidation. Consolidation involves the aggregation of data, i.e. the total number of students at the university, total courses, and average GPA;

2. *Drill-down.* This is the opposite of consolidation that involves more details, i.e. Example is the break down of the total number of students onto the different nationalities that belong to different majors with different GPA on average;
3. *Slicing and dicing.* Slicing and dicing refers to the ability to look at the database from different viewpoints.

5-4 Association rules

Association rules are always defined on binary attributes. This sort of attributes makes it easy to describe the student profiles in our example database. Example is:

Male student, Nautical department ⟶ *Egyptian nationality*

This means that if the student 's gender is male and the major is nautical, so the student's nationality is Egyptian. And this happening 65% confidence level.

In fact, the number on possible association rules that might be found in our database-and the like- is almost infinite. However, there is no algorithm that will automatically give us everything of interest in the database. Some roles are found to be useless (Argawal, et al., 1996). For example: the students that live in Miami tend to have the Hotels major, or the relationship between odd or even registration numbers and GPA achieved in a certain major, and the like rules. One might ask if it is easy to find such a rule in data mining of this type of relationship. The answer is "YES". Many techniques exist to find such rules (Adriaans and Zantinge, 1996). The level of *Confidence* is another aspect that worth mentioning in this context. Confidence is the percentage of records that holds true for the fact under studying within the group of records.

5-5 Decision trees and rules

A decision tree is a predictive model that can be viewed as a tree. Each branch of the tree is a classification question, and the leaves are partitions of the data set with their classification (Berson and Smith, 1997). Another definition for the decision tree uses logical methods of describing regions of state. These logical methods could be interpreted in a "IF..THEN" rules space (Pyle, 1999). One variable is studied individually in a decision tree. The start point is found when the variable that best classifies the state space is determined and consequently the true state is also raised. The algorithm of the tree then looks for another classifying variable and another splitting rule. This process continues until some ending criteria is found. Another algorithm for developing the decision tree is called ID3 is iterative algorithm. The ID3 start with a subset of the data called *window*. The window is chosen at random and then develops the tree which is correctly classifies the subset into branches. Then all other data are classified using the tree. If each data record finds its classification so the process terminates. If not, a selection of the incorrectly classified data records is added to the window and the process continues (Quinlan, 1986).

The decision tree is a useful technique in both data mining and predictive modelling processes. It prevents the problems of overfitting (it happens when the algorithm searches in the limited data sets, so the algorithm might overfit the data) and handling of missing data that majority of the mining techniques always leave these problem in the user side. The decision tree classifies the data into branches without loosing any of the data records. Another fact about the decision trees states that the tree works well with the relational database management systems with high accuracy (Berson and Smith, 1997).

5-6 Artificial Neural Networks (ANN)

The term neural network refers to artificial neural networks (ANN), because the true neural network is the biological systems (our brains) that can make predictions detect patterns and

learn. An ANN is a computer programs that implements complex pattern detection and machine learning algorithms to build predictive models from large database(s). In order for the ANN to detect patterns in the data sets, it should learn to detect these patterns and make predictions, is the same like a human being does. ANN are widely used in many business applications. ANN are also of different types; they can be used for clustering creation. Most of the ANN are hard for users to understand because they lack clarity, however, vendors tend to introduce the ANN in visualization formats that make them understandable. Also an ANN is a time consuming process to build, however, it is a very powerful predictive technique that require some data preprocessing, good understanding of the problem and the target of prediction. It also requires the setting of some parameters that will drive its mission.
The ANN has both advantages and disadvantages, the big advantage is the highly accurate predictive accuracy that could be applied to different problems. The big disadvantages are the lack of ease of use and the difficulties of deployment (Berson and Smith, 1997).

5-7 Nearest neighbour

The nearest neighbour technique is one of the oldest techniques used in data mining. Nearest neighbour is a prediction technique that is used to perform data *clustering*. Many other techniques are used for clustering *like the complete linkage method or the furthest neighbour, Centroid cluster analysis, Media analysis, Group average method,* and *automatic interaction detection* (Everitt, 1981). Since they all perform the same job, which is data classification or data clustering, one of them is illustrated here.
It begins with a distance *matrix* that shows the individual groups and end with a *dendrogram* showing the successive fusions which culminates all individuals in one group (Everitt, 1981). Groups of single individuals are fused with their nearest members. Each fusion decreases by one the number of groups. The concept of nearest neighbour states that records that are close to each other live in each other's neighbour. Assume that we want to predict the performance of a group of students and we have the complete DB of students. The basic idea is that students that are of the same group or cluster will show the same performance. So if we need to predict the performance of a student we first look at the students that are lose to him in the database.

Nearest neighbour is used in many business applications like bankruptcy prediction in the banking industry and handwriting recognition. The nearest neighbour technique is automated but require some preprocessing of data in converting some predictions into values that will be used to measure distance. However, unordered variables like eye-color should be transformed into the distance between each other when there is a match found (Berson and Smith, 1997). The techniques will give the user the high-level view of what is going in the database. The nearest neighbour is optimized for prediction of new records rather than exhaustive extraction of interesting rules from the database. Most of the text retrieval systems are built around the nearest neighbour technologies. The nearest neighbour takes into account all the predictors into consideration that is helpful for prediction but makes the model a little bit complex so it can not be described in a rule format.

A problem in the nearest neighbour is the existence of tables consists of high number of attributes. Nearest neighbour does not perform well if there is a very big number of independent attributes because this will need a multi-dimensional search space. Possible problems associated are millions of very complex and it is also possible to find an equal space between data that is hard to define the nearest of them. One approach here is to limit the number of attributes by finding the relative importance of them and work on them, or to search for another technique that is able to handle the large number of attributes.

5-8 Genetic Algorithms

The term is a combination of both biology and computer disciplines, and sometimes referred to as simulated evolution. Berson and Smith (1997) said "Genetic algorithms loosely refer to these simulated evolutionary systems, but more precisely these are the algorithms that dictate how populations of organisms should be formed, evaluated and modified". Genetic algorithms are used to create the biological evolution version of computers. They start on a small program and then mature themselves as the human organisms undergoing natural evolution. Over time, these programs on the computer improve in their performance and as a result increase the efficiency of resolving a certain problem. In many ways genetic algorithms are close to the biological evolution, the analogy is like the following table (1):

Biology	**Genetic Algorithms**
-Organism	Which is the computer program being optimized
-Population	The collection of programs undergoing simulated evolution
-Chromosome	The chromosome encodes the computer program
-Fitness	The calculation with which a program value is determined for survival of the fittest
-Gene	The basic building block of the chromosome that defines one particular feature of the simulated organism
-Locus	The location of the chromosome that contains a specific gene
-Allele	The value of the gene
-Mutation	The random change of the value of the gene
-Mating	The process by which two simulated programs swap pieces of them in a simulated crossover
-Selection	The best program is retained and the less successful are excluded by deleting them from computer memory.

Table (1). Genetic Algorithms and Biology.

Genetic algorithms are used to find optimal clusters based on a defined profit measure. Genetic algorithms by themselves do not detect outliers and do not create rules. They have been used to optimize the nearest neighbour classification systems for predicting sequences in time series (Berson and Smith, 1997).

5-9 Probabilistic graphical dependency technique

These models specify the probabilistic dependencies, which underlie a particular model using a graphical structure. The model specifies which variables are dependent on each other. These models are used for categorical or discrete-valued variables, however, some extensions allowed the use for real-valued variables also. Within artificial intelligence and statistics these models are initially built in the context of the probabilistic expert systems (ES). Although, graphical models induction is still not a mature discipline, it is of interest to the KDD applications since graphical forms of the model are easily understood by users (Fayyad et al., 1996).

6- Example for the KDD process applied to student recruitment systems

The KDD process will be applied to a sample data records (1600 records) extracted from the students' recruitment system database used at AASTMT[3]. Not all of the database fields were used for this purpose; rather the data, which are of importance to the executive managers that,

[3] Data used with permission.

are entailed in the students' recruitment system functions. Questions that might be of value are the following:

-How many students applied? And of them how many have got acceptance?

-What is the nationality distribution of applicants?

-What is the minimum total marks accepted at the AASTMT? Also, what is the minimum percentage accepted at each department?

-How to visualize the nationalities, age, and departments in one 3-d format?

-How to predict the student profiles?

-To what extent these numbers reflect the AASTMT competitive position?

-What do the competitors' statistics say?

-What are the countries that need marketing?

6-1 Data selection

In the sample under study we started with 1600 records of the applicants, from which a number of 1100 has been accepted at AASTMT (AASTMT application records, 1995; AASTMT statistics, 1995). The records consist of serial number, application number, first name, nationality, sex, address, desire(s), percentage grade, accepted/ rejected, and department. In order to facilitate the KDD process a copy of this operational data is drawn and stored in a scratch database file[4], a sample records of this database is given here in table (2).

SN	App.No	F.Name	Sex	Nation	City	Desire	% mark	A/R	Dept
1	697	Mary	F	Egy	Alex	Hot	95	A	Hot
4	1079	Asser	M	Egy	Alex	Bus	70	A	Hot
197	1484	Ismael	M	Syr	Dam	Eng	67	A	Eng
1	1570	Mohamed	M	Egy	Alex	Hot	90	R	
4	19	Lamees	F	Pal	Dub	Bus	83	R	

Table (2). Sample of the original data.

Note: SN stands for serial number, App.No for application number, F.Name for first name, Nation for nationality, A for accepted & R for rejected, and dept for department. Also, the departments, and nationality names have been curtailed for space purpose in the table.

6-2 Cleaning

Several methods are available to clean the data i.e. remove errors. Some of these methods can be executed in advance while others are only invoked after errors are detected at the coding or the discovery stage (Adriaans & Zantinge,1996). A very important element in a cleaning operation is the *de-duplication* of records (table 3). In the student database file one student may be presented by more than one record. For example, if a student applied from abroad and then a relative also applied on behalf of the student, this should be positioned another application for the same person, this would be clear from the check of data. Another source of error is that students change their address without notifying the Admission & Registration office. There are also cases in which people spell their names incorrectly or give incorrect information about themselves by slightly misspelling their name or by giving a false address. Data cleaning processes affect the quality of the mining process, because as much seriously the process of data cleaning was performed, the results of data mining would be helpful and trustworthy.

SN	App.No	Name	Sex	Nation	City	Desire	% mark	A/R	Dept
270	316	Ehab	M	Egy	Alex	Eng	70	A	Eng
271	623	Ehab	M	Egy	Alex	Eng	70	A	Eng

Table (3). De-duplication of records.

[4] MS-Access was used by the researcher for this purpose.

In the present example we have two different records for the same student data, so this may be due to two persons submitted the same student data without the student being aware of that. Of course, we can never be sure of this, but de-duplication algorithm using analysis techniques could identity the situation and present it to a user to make a decision. The second type of data errors that frequently occurs *is the lack of domain consistency*, (table 4).

SN	App.No	Name	Sex	Nation	City	Desire	% mark	A/R	Dept
541	350	Ahmed	m	Egy	Cairo	Eng	70	A	
215	476	Ahmed	m	Alex	Lyb	Eng	70	A	Eng

Table (4). Domain consistency.

Notice that in table (4) the first student department is empty, however, this attribute should have a value. Empty values are a source of problems to the data mining process, because this might affect the type of patters discovered. If the data item is not defined it should be NULL. On the second record there is inconsistency in the domain of nationalities and cities, the opposite is true; replacement would happen between city, and nation. The result is shown in table (5).

SN	App.No	Name	Sex	Nation	City	Desire	% mark	A/R	Dept
541	350	Ahmed	m	Egy	Cairo	Eng	70	A	NULL
215	476	Ahmed	m	Lyb	Alex	Eng	70	A	Eng

Table (5). Domain consistency-1.

Needless to say this is disastrous in a data mining context. Since if information is unknown it should be represented as such in the database. In our example, we have replaced part of the data with NULL values and corrected other domain inconsistencies.

6-3 Enrichment

In our present example, assume that we have got some extra information about the students' family annual income, and the student secondary school, where it is possible to get or buying them. Table (6).

SN	App.No	Name	Income	School	Nation	Dept
100	801	Ola	200.000	Saudi Sch.	Sau	NULL
200	802	Alaa	350000	IGCSE	Leb	Eng

Table (6). Enrichment.

6-4 Coding (Pre-coded data)

In the next stage, we select only those records that have enough information to be of value. Notice that the extra information should be added to the original data. Table (7).

SN	App.No	Name	Inc.	Car	Sex	Nation	City	Desire	%	A/R	Dept
911	350	Tarek	Null	Null	M	Egy	cairo	Eng	60	A	Null
111	714	Wael	150	Null	M	Null	Alex	Mar	85	A	Eng

Table (7). Enriched table.

A general rule states that any deletion of data must be a conscious decision, after a thorough analysis of the possible consequences. However, in some cases lack of information can be a valuable indication of interesting patterns. In the presented cases in table (7) for the students Tarek, and Wael, we lack some vital data concerning them, so we choose to exclude their records from the final sample. Of course, this decision is questionable, because there may be a causal connection; however, it is better to delete incomplete information instead of getting incorrect results (Adriaans and Zantnige, 1996). Next we carry out a projection of the records. In our example we are not interested in the students' name, so their names are removed from the sample. Up to this point, the coding phase consisted of nothing more than simple SQL operations but now we are entering the stage where we will be able to perform some data transformations. By this time, the information in our database is much too detailed to be used

as input for pattern recognition algorithms. For example, department, nationalities, and cities represent a complex set of data that has to be coded before used as inputs.

6-4-1 Coding (post-coding data)

Coding, therefore is a creative activity that has to be performed repeatedly in order to get the best results. Income even is a source of problem. One solution might be to transform income into categories and describe each category's characteristics. For example: annual income between 300.000-200.000 L.E is called moderate-and coded 01, and annual income between 1,000,000-2,000,000 L.E is called premium income-and coded 05. *So that we can obtain information such as the following:*

1. Students applied from nationality Egyptian, in the Maritime department, have a premium income group. OR
2. Students from Syria, always have Engineering as a major, with average marks of 60%.

Instead of doing a student data analysis, the relationships between students of different nationalities are important. This means that we will not be investigating the connections between individual attributes, but between different student profiles. Before going in depth to the results, a complete coding process should take place.

SN	App.No	Inc	Cat	Sex	Nation	City	Desire	%	A/R	Dept
911	350	05	Null	m	10	101	1	60	A	1
111	714	01	Null	m	30	301	4	85	A	4

Table (8). The coding effect.

Table (8) represents the new table that results from the coding process. A table in this format is not very helpful if we need to find relationships between different student profiles. Each student is represented by one record. That is, instead of having one attribute with five different possible values, we create twenty binary (*0 and 1 are the components of the code*) attributes, one for each nationality. If the value of the nationality attribute is "10" this means that the student is Egyptian, 20 means Sudanese, 30 means Libyan etc. Such an operation is called (*Flattening)* an attribute with *cardinality n* is replaced by *n* binary attributes. Applying the concept Flattening to the Sex results in 01, 02, which means that male is replaced by 01, and female is replaced by 02 (*cardinality 2)*. This is a coding operation that occurs frequently in a KDD context (Adriaans & Zantinge, 1996).

6-5 Data Mining Techniques

The data mining techniques were discussed with respect to many classification mechanisms, however, the KDD example here will not definitely handle all of the techniques because there is no single application that uses all of the data mining techniques, and that is due to the differences between techniques and data sets and the application domains.

6-5-1 Traditional query tools

A good way to start is to extract some simple statistical information from the data set. *Table (9) provides statistics on students.*

Department	Private*	Sponsor**	Transfer	Total
Nautical	54	6	0	**60**
Maritime Eng.	37	2	0	**39**
Mechanical Eng.	51	0	2	**53**
Computer Eng.	93	6	6	**105**
Power Eng.	64	0	0	**64**
Electronics Eng.	78	5	5	**88**
Construction Eng.	61	0	0	**61**
Managerial Eng.	58	0	0	**58**
Business Adm.	181	6	7	**194**

Hotels & tourism	27	0	1	**28**
Grand total	**704**	**25**	**21**	**750**

Table (9). Statistics.

***Private** private student means that he/she pays for himself/herself.

****Sponsor** sponsored student appears when a certain agency pays for the student fees and accommodation, regardless of the reason.

A number of 350 students should be added to the grand total of the previous table"1100 total accepted", they represent the number of students that will join the preparatory program so that, after succeeding they can join the next semester unless rejected. From table (9), and the original data sets, a lot of shallow knowledge can be extracted and represented. We can see that the percentage of sponsored students at the AASTMT-according to the sample data records- is about 3%, as well as the transfer students "*if the AASTMT has 1000 students 940 are predicted to be private, 30 are transfer, and 30 are sponsored*". The analysis is not a goal in itself, rather the implications of the analysis to the executive who will take a decision according to that analysis. For example, the drill down capability reveals that the 3% of students who are sponsored historically joined the Nautical department which has a little demand nowadays (AASTMT statistics, 1990-1997). May be the reason is the student is actually sponsored but he claims the opposite to save money, this is because sponsored students pay higher fees.

Data mining represents the output results to the executives so that they can benefit from and go further in analysis and drill-down capability. For instance, the fact that the Business Administration has the highest demand may be traced and mined as follows:

-Maybe there is a trend in the labor market to give the business graduates greater number of work opportunities.

-Maybe the competitor universities are weak in this field.

-Maybe the excellent facilities the AASTMT has made this possible.

-Maybe the staff members of the Business Administration department are superior to its rivals, and use nice textbooks.

-Maybe the marketing efforts of the AASTMT are biased to the Business Administration department.

-Maybe the reason of this is that the Business Administration graduates excel at their work.

Regardless of what the reason is, the drill-down capability is an important tool for executives, and this is the core benefit the data mining offers to the Executive managers. Data mining can make it easy to the executive to analyze the results. The graph shows that some of the Engineering departments (Marine, Power, Construction, Managerial) and the Nautical have a 0% transfer, so this might be due to the following:

-Students of the other departments see no value of applying to these departments;

-Course structure makes it difficult for students to transfer to them;

-A decrease on student GPA if transferred to them, due to the cancellation of some courses he has completed that will not be counted;

-These departments do not accept transfer students.

No matter what the reason is the analytical ability offer the different possibilities to the executives and find the hidden knowledge and patterns, then the role of the executive is to take corrective actions.

6-5-2 Visualization techniques

In the next figure (3), visualization technique plays a great role, that is it puts the information we have in an easy to understand way. Table (9) has been transformed to the following figure (3).

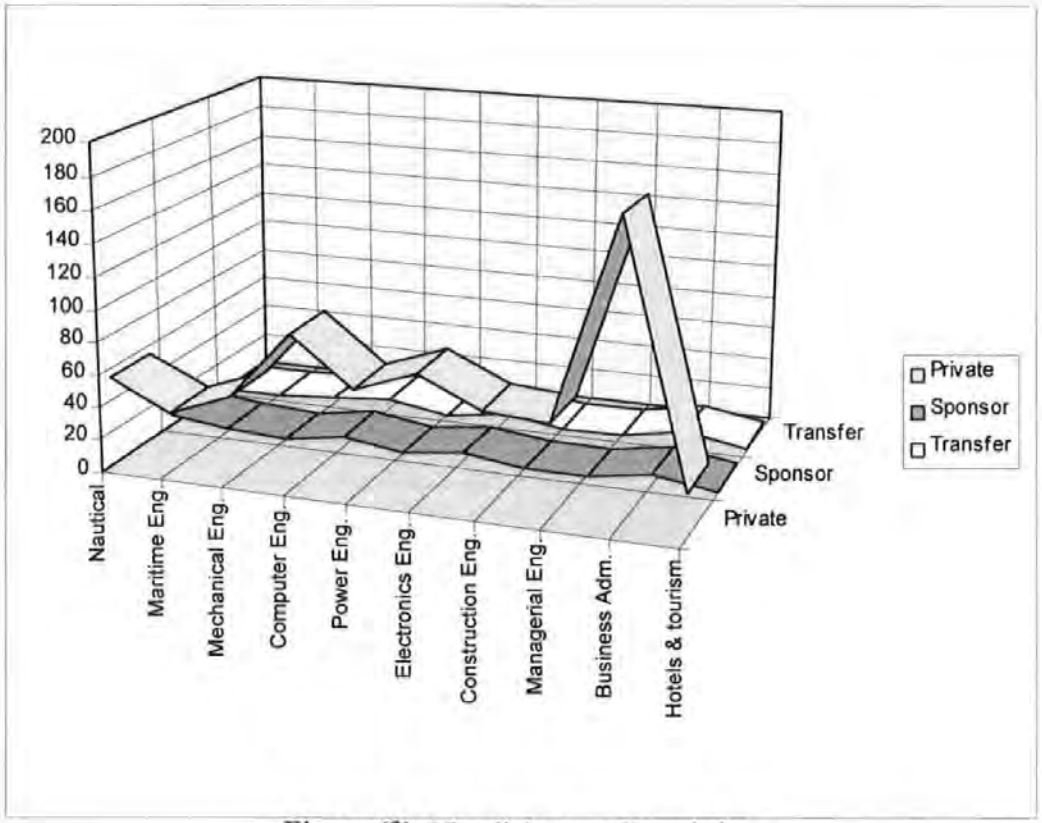


Figure (3). Visualizing new Knowledge.

6-5-3 OLAP

We can plot table (9) in the following format in figure (5). The figure represents the students' data against departments, regarding if they are sponsored, private, or transfer students, but in a category of details level.

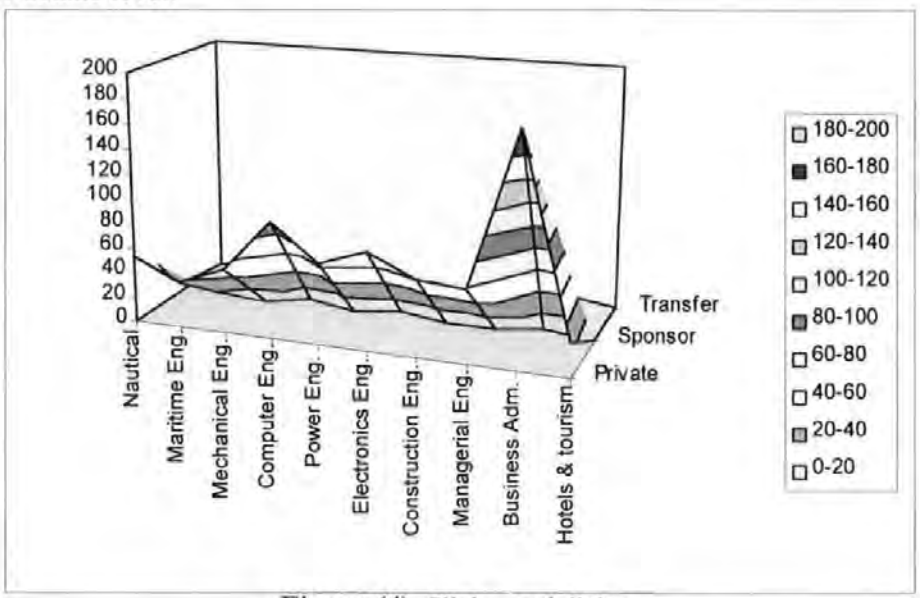


Figure (4). Slicing and dicing.

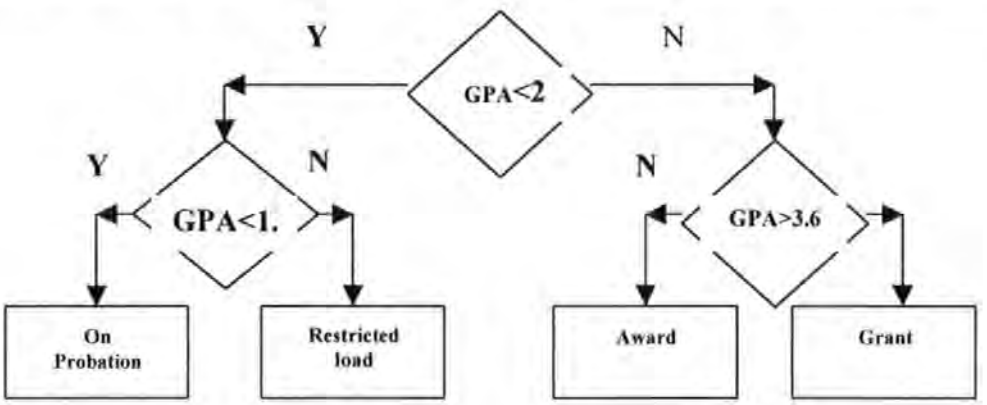
6-5-4 Association rules

For the purpose of our database set we can develop a set of rules. For example:

93% of the students taking the Bus. Adm. major ⟶ 25% of total students.
The previous rule importance is that, what is applied to the private students at the Business Administration department is applied to 25% of the total private students.

6-5-5 Decision trees and rules

A decision tree like the following can be found in the admission and registration function.



- The example follows the credit hours system at the AASTMT.

Rule 1: IF GPA is less than 1.6 THEN student is on probation.
Rule 2: IF 1.6 is less than or equal GPA is less than 2THEN student has restricted load.
Rule 3: IF 2 is less than or equal GPA is less than 3.6 THEN student takes award.
Rule 4: IF 3.6 is less than or equal GPA is less than or equal 4 THEN student takes grant.

Figure (5). Decision tree for students'grants.

6-5-6 Nearest neighbour

Assume that we have 4 student GPA groups. As in the following matrix:
Assume that we have 4 student GPA groups. As in the following matrix:

Matrix G1=	**1**	**2**	**3**	**4**
1	0	3.2	3.4	3.8
2	3.2	0	4	2
3	3.4	4	0	1.9
4	3.8	2	1.9	0

Step (1): The smallest entry is G34, so group 3 and 4 are fused together. Distance between G34 and groups 1 and 2: G(34)1=min {G31,G41}=G31=3.4
G(34)2=min {G32,G42}=G42=2

Step (2): Matrix G2 is the following.

Matrix G2=	**1**	**2**	**(34)**
1	0	3.2	3.4
2	3.2	0	2
(34)	3.4	2	0

The smallest entry is G2(34), so group 2 and (34) are fused together. Distance between G2(34) and group 1: G(34) 1=min {G31,G41}=G31=3.4

Matrix G3=	**1**	**2(34)**
1	0	3.2
2(34)	3.2	0

The dendrogram is the following.

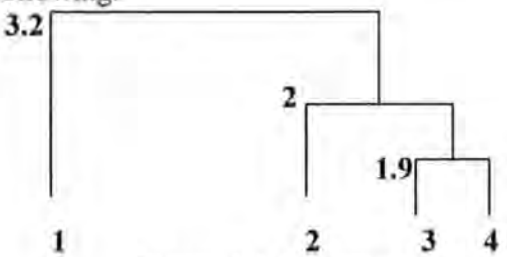


Figure (6). Dendrogram.

6-6 Reporting

Reporting the results of data mining can take many forms. The illustration already described in this chapter has given a good idea of the options available. *In general, one can use any report writer or graphical tool to make the results of the process accessible* (Adriaans and Zantinge, 1996). The material in this chapter has also provided a good indication of the interactive character of the KDD process: there is a consistent interplay between the selection of data, cleaning, data mining, and the reporting results.

7- KDD future research

1. *Larger databases/algorithm scalability.* The mining techniques as step in the KDD process should be able to handle database with hundreds of tables and fields, as well as with millions of records. These databases normally are of gigabytes and sometimes of terabytes in size. The techniques should be efficient in dealing with these volumes of data;
2. *High dimensionality.* This problem is closely related to the very large number of fields/variables in the database, so the dimensionality is high. The high dimensionality creates a larger search space, which causes the model induction to be complex and time-consuming. This also will increase the probability that the data mining technique will find spurious patterns;
3. *Overfitting.* This happens when the mining algorithm searches for the best parameter for one model using a limited data set, which may lead to an overfit of the data and models resulting poor performance;
4. *Assessing statistical significance.* This problem is very close to the overfitting one, it happens when the system is searching over many possible models. If the system is testing N models at the 0.001 significance level, then on average, with random data, N/1000 of these models will be accepted as significant;
5. *Changing data and knowledge.* Rapidly changing data may cause the discovered patterns to be irrelevant and spurious. Example is the stock market data;
6. *Missing and noisy data.* Here the problem emerges when important data items are missing. This problem often occurs when the database design is carried out without taking into consideration the discovery process;
7. *Complex relationships between fields.* Hierarchically attributes or relations between attributes require the mining algorithm to utilize them. Traditionally mining techniques have been developed to deal with simple attribute-value records;
8. *Understandability of discovered knowledge.* One critical factor in KDD applications is to make the output of the systems human-understandable;
9. *Data security.* This is related to the security of the original data and how should the mining techniques deal with this data without violating their security.

-Conclusion

KDD is the process of fining hidden knowledge, patterns and unknown facts from the data sets. It is a multi-disciplinary field. Data mining is a step in the KDD process. Goals of the data mining techniques are prediction and description. The data mining tasks are clustering, classification, summarization, dependency, regression, and change detection. Data mining techniques are many. The techniques used for different tasks up to the goal of the KDD process. The application of the KDD process will be of value at the universities, where it will introduce knowledge to the decision-makers to enable them to reach a better quality of decisions.

- References

Arab Academy for Science and Technology & Maritime Transport (AASTMT) statistics. Different volumes from 1990-1997.

Adriaans, P., and Dolf Zantinge. (1996). Data Mining, Addison Wesley Longman.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. (1996). Fast discovery of association rules. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press.

Agrawal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, and Sunita Sarawagi. (1996). On the Computation of Multidimensional Aggregates. In Proceedings of the 22nd VLDB Conference Mumbay, India, 1996.

Anand, S., Bell, D., and J. Hughes. (1995). Evidence based discovery of knowledge in databases. In Colloquium on Knowledge Discovery in Databases, IEE, UK, pp. 9/1-9/4.

Barquin, R. (1997). A Data Warehousing Manifesto. Planning and Designing The Data Warehouse, Prentice Hall, pp.3-16.

Barquin, R., Paller, A., and Herb Edelstein (1997). Ten mistakes to avoid for data warehousing managers. Planning and Designing The Data Warehouse, Prentice Hall, pp.145-156.

Barrow, C. (1992). Implementing an Executive Information System: Seven steps for success. In Watson, H., Rainer, R., and Goudeshel, G. Executive Information Systems: Emergence, Development, Impact. John WILLEY and sons Inc, pp.107-116.

Bennett, J. (1983). Building decision support systems, Addison-Wesley.

Berson, A. (1996). Client/Server Architecture, McGraw-Hill.

Berson, A., and Stephen Smith. (1997). Data Warehousing, Data Mining, & OLAP, McGraw-Hill.

Brachman, R., and Tej Anand. (1996). The process of Knowledge Discovery in databases. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.37-57.

Buntine, W. (1996). Graphical Models for Discovering Knowledge. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.59-81.

Canavos, G., and Don M. Miller, (1995). Modern Business Statistics, Duxbury.

Chen, M., Han, J., and Philip S. Yu. (1996). Data Mining: and overview from a database perspective. IEEE Knowledge and Data Engineering, (8:6)pp.866-883.

Chengular-Smith, I., Ballou, D., and Harold L. Pazer. (1999). The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. IEEE Transactions on Knowledge and Data Engineering, Nov-Dec 1999, Vol.11, number 6, pp.853-864.

Clifton, C., and Marks, d. (1996). Security and privacy implications of data mining. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.

Daniel, A., and Kriegel, H. (1996). Visualization techniques for Mining large databases: A Comparison. IEEE Knowledge and Data Engineering, (8:6).pp. 923-938.

Date, C. (1995). An Introduction to Database Systems, Addison-Wesley.

David, W., Vitcent, T., Ada, W., and Yongjian, Fu (1996). Efficient Mining of Association Rules in Distributed Database. IEEE Knowledge and Data Engineering, (8:6) pp. 911-922.

Devlin, B. (1997). Data Warehouse: From Architecture to Implementation, Addison-Wesley.

Dzeroski, S. (1996). Inductive Logic Programming and Knowledge Discovery in Database. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press pp. 117-152.

Edelstein, H. (1997). An introduction to data warehousing. Planning and Designing The Data Warehouse, Prentice Hall, pp.31-50.

Elam, J., Jarvenpaa, S., and David A. Schkade. (1992). Behavioral Decision Theory and DSS: New Opportunities for Collaborative Research. In Stor, E., and Benn R. Konsynski. Information Systems and Decision Making, IEEE, pp.51-74.

Everitt, R. (1981). Cluster analysis, Halsted press.

Fayyad, U., Piatetsky, G., and Padharic Smyth, (1996). From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.1-34.

Gaines, B. (1996). Transforming Rules and Trees. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.205-220.

Ganti, V., Gehrke, J., and Raghu Ramakrishnan (1999). Mining very large databases. IEEE Computer magazine, Aug 1999,pp.37-45.

Garcia-Molina, H., Labio, W., and Jun Yang. (1998). Expiring Data in a Warehouse. In the Proceedings of the 24th VLDB Conference, New York, USA, 1998.

Gray, P, Alter, S., DeSanctis, G., Dickson, G., Johansen, R., Kraemer, K., Olfman, L. and Douglas R. Vogel. (1992). Group Decision Support systems. In Stor, E., and Benn R. Konsynski. Information Systems and Decision Making, IEEE, pp.75-136.

Gray, P., and Hugh J. Watson. (---). The New DSS: Data warehouse, OLAP, MDD, and KDD, ---.

Guyon, I., Matic, N., and Vladimir Vapnik. (1996). Discovering Informative Patterns and Data Cleaning. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.181-203

Hadden, E. (1998). Building for Successful Data Warehouses and Data Marts, BBS International.

Hadden, E. (1998). Planning for Successful Data Warehouses and Data Marts, BBS International.

Han, J., Fu, Y., Wang, W., Krzysztof, K., and Zaiane, O. (1996).DMQL: A data mining query language for relational databases. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.

Han, J., Lakshmanan, L., and Raymond T.Ng. (1999). Constraint-based,

multidimensional Data mining. IEEE Computer magazine, Aug 1999,pp.46-50.
Imielinski, T. (1996). From file mining to database mining. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.
Inmon, W.(1993). Building the Data Warehouse. QED publishing Group.
Inmon, W., and Richard D. Hackathorn. (1994). Using The Data Warehouse, John Wiley & Sons.
Jaeger, M., Mannila, H., and Weydert, E. (1996). Data mining as selective theory extraction in probabilistic logic. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.
Jajodia, S., and Ravi Sandhu. (1991). A Novel Decomposition of Multilevel Relations Inot Single-Level Relations. In proceedings of IEEE Symposium on Security and privacy, Oakland California, May 1991, pp300-313.
Khosla, I., Kuhn, B., and Soparkar, N. (1996). Database search using information mining. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.
Klein, M., and Leif B. Methlie. (1995). Knowledge-Based Decision Support Systems, John Wiley & Sons.
Livingston, G., and Bob Rumsby, (1997). Database Design for The Data Warehouses: The Basic Requirements. Planning and Designing The Data Warehouse, Prentice Hall, pp.179-198.
Mattison, R. (1997). Data warehousing and data mining for telecommunications, Artech House.
Milne, R., and Nelson, C. (1995). Knowledge Guided Data Mining. In Colloquium on Knowledge Discovery in Databases, IEE, UK, pp. 6/1-6/3.
Mimno, P. (1997). Data Warehousing Architectures. Planning and Designing The Data Warehouse, Prentice Hall, pp.159-177.
Munton, A., Silvester, J., Stratton, P., and Helga Hanks. (1999). A Practical Approach to Coding Qualitative Data, WILEY.
Paller, A. (1997). A Roadmap To Data Warehousing. Planning and Designing The Data Warehouse, Prentice Hall, pp.17-29.
Palvia, P., Kumar, A., Kumar, N., and Hendon, R. (1996). Information Requirements of a global EIS: An exploratory macro assessment. Decision Support Systems, 16:169-179.
Pyle, D. (1999). Data preparation for Data Mining, Morgan Kaufmann Publishers.
Quinlan, J. (1986). Induction of Decision Trees. Machine Learning, (1) pp.81-106.
Raden, N. (1997). Choosing The Right OLAP Technology. . Planning and Designing The Data Warehouse, Prentice Hall, pp.199-224.
Ramakrishnan, N. and Ananth Grama. (1999). Data mining: From serendipity to science. IEEE Computer magazine, Aug 1999,pp.34-37.
Saraee, M., and B. Theodoulidis, (1995). Knowledge Discovery in Temporal Databases. In Colloquium on Knowledge Discovery in Databases, IEE, UK, pp. 1/1-1/4.
Silberschatz, A., and Tuzhilin, A. (1996). User Assisted Knowledge Discovery: how much should the user be involved. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.
Silver, M. (1991). Systems That Support Decision Makers: Description and Analysis. John Wiley & Sons, New York.
Taha, Y., Helal, A., and Ahmed, K. (1997). Data Warehousing: Usage, Architecture, and Research Issues. ISMM Microcomputer Application Journal,16(2):1-8.
Thearling, K. (1999). Increasing customer value by integrating Data mining and campaign management Software. Direct Marketing Magazine, Feb 99.

Turban, E. (1993). Decision Support and Expert Systems, Management Support Systems, Macmillan.

Turban, E., and Jaye Aronson, (1998). Decision Support Systems and Intelligent Systems, Prentice Hall.

Widom, J. (1995). Research problems in data Warehousing. In proceeding of the 1995 ACM CIKM, pp. 25-30.

Widom, J. (1995). Research Problems in Data Warehousing. In the Proceedings of 4th International Conference on Information and Knowledge Management (CIKM), Nov. 1995.

Zhang, X., and Elke A. Rundensteiner. (1998). Data Warehouse Maintenance Under Concurrent Schema and Data Updates. Computer Science Technical Report Series, Aug 1998, pp. 1-29.

DEVELOPING STAR SCHEMA STRUCTURES- A PRACTICAL STUDY[5]

Ahmed El-Ragal
University of Plymouth
Business School, Management dept
Drake Circus, Plymouth, PL4 8AA
Devon, UK
Tel: +44-01752-232850
Fax: +44-01752-232853
E-mail: a.el-ragal@plymouth.ac.uk

Terry Mangles
University of Plymouth
Business School, Management dept
Drake Circus, Plymouth, PL4 8AA
Devon, UK
Tel: +44-01752-232852
Fax: +44-01752-232853
E-mail: terry.mangles@pbs.plym.ac.uk

ABSTRACT

This paper investigates many of the practical issues surrounding the development and implementation of a star schema structured data warehouse. The paper introduces and summarizes the organizational requirements that are required to underpin the student recruitment process in higher education. These requirements have been identified following an in-depth survey of the recruitment process in Egyptian Universities. This survey was used to identify the required data sources together with the likely users and their information needs. The survey was sent to senior managers within the Egyptian Universities (both private and public) with responsibility for student recruitment, in particular the admission and registration processes. Further, access to a large database has allowed us to test the practical suitability of using a data warehouse structure and knowledge management tools within the decision-making framework. The design of the proposed data warehouse will be developed and illustrated using CASE tools. It is not the intention of this paper to compare tools but to illustrate the use and benefits that these can accrue to the systems developer. In particular, the benefits of matrix verification within the data warehouse design process will be explored and the paper will illustrate how these tools can be used to produce efficient and bug free data warehouse. Finally, the paper will discuss the use of front-end tools to develop an easy to use system without which the data warehouse would not be used.

[5] The 4th SAIS 2001 Conference, in Savannah-GA, USA, 2-3 March 2001, pp. 116-136.

DEVELOPING STAR SCHEMA STRUCTURES- A PRACTICAL STUDY

INTRODUCTION

To survive and succeed in today's global environment organizations' needs from data have become more variable because of the following (Srivastava and Chen, 1999; Barquin, 1997; Paller, 1997); decisions need to be taken quickly and correctly using all the available data, users are not computer professionals (Berson and Smith, 1997), the amount of data is increasing, it is becoming important to be able to obtain a comprehensive and integrated view of the enterprise, decisions sometimes require historical analysis, and identifying trends in the business (Onder and Nash, 1999).

DATA SOURCES

To respond to these needs various departments in organizations store data about internal transactions and about their external environment. Further, organizations need to store these data for a number of years in a historical (archival) database to trace patterns in this history and/or to meet legislative requirements e.g. tax regulations. Decision makers to reach a better business understanding and improve the decision quality then access these data sources. Sometimes decision makers keep their own experience in a separate database (Turban and Aronson, 1997). These different types of data are illustrated bellow.

Internal data

These are the data sources of an organization that cover the whole business, e.g. data about employees, daily transactions, products, stock levels, customers etc. These internal data are stored in one or more databases. Internal data sources are the output of using transaction processing systems (TPS) in organizations, because these are the systems that store data about business transactions. All these organization internal data sources are sometimes referred to as operational data sources (ODS) (Hadden, 1998) or as the on line transaction processing systems (OLTP) (Berson and Smith, 1997; Devlin, 1997);

External data

This is data that comes to the organization from outside sources. There are many types e.g. government reports, federal publications, research institutions, commercial data banks, access to suppliers and customers' databases, and the Internet (Turban and Aronson, 1997). External data sources are used in EIS and DSS to enhance the strategic and long-term decisions. Examples for the commercial data banks include CompuServe, Compustat, Data Stream, Dow Jones Information service, and Lockheed Information systems;

Archival or historical data

When an organization needs to store data about a specific topic for several years, e.g. 5 years, it uses an archival or historical database. The archival database can contain either internal or external data sources or both.

Personal data

This data source includes the manager's own experience and opinions and/or estimates about market share, additional customer data or other policies. These personal data sources are used in EIS and DSS.

DATA STRUCTURES

As the previous studies have shown organizations have an increasing need for storing and processing data to meet the organization's information need. However, at the operational level data is collected into a number of separate data sources which will form the basis of the ODS and the data warehouse (DW).

For organizations to be able to develop operational data stores for their different data sources there are various database structures that can be used. There are three fundamental DB structures: relational, hierarchical, and network. There are also new DB structures e.g. object-oriented, multi-media, and the star schema structure (Livingston and Rumsby, 1997; Date, 1995; Elmasri and Navathe, 1994; Pratt and Adamski, 1987). These different types of database structures are discussed in the following sections.

The relational structure

This is the most frequently used database structure within the DBMS context (Date, 1995; Elmasri and Navathe, 1994). It is also the dominant database structure in DSS applications, and frequently used in the development of a DW (Turban and Aronson, 1998, Berson and Smith, 1997).

Hierarchical

A hierarchy (or tree) is a network in which nodes are connected by links such that all links point in the direction from child to parent. Each node has one parent and there is always one path between any two nodes (Parsaye, et

al., 1989). The hierarchical model stores the data fields in a top-down order. The database looks like a tree, and there are links between related fields.

Network

The network structure sometimes called the CODASYL structure, uses additional pointers to give the hierarchical structure more flexibility. The network structure allows more complex links between nodes. Thus the hierarchical structure may be viewed as a special case of the network structure where each node is linked to a parent node only (Parsaye, et al., 1989). The network structure saves storage space through the sharing of data items. In the network structure only the one-to-one and many-to-one links are allowed.

Object-oriented

This structure is used with complex applications which require accessibility to data that have complex and inter-related relationships (Date, 1995). For example computer aided design and manufacturing (CAD/CAM), computer integrated manufacturing (CIM), and geographic information systems (GIS). Relational, hierarchical, or network data structures can not support these applications efficiently.

Multi-media

The multi-media structure manages data in many formats; text, numeric, images, bit-maps, pictures, hypertext, video clips, sounds and multi-dimensional images (virtual reality).

DATA WAREHOUSE (DW)

Several DW definitions have been proposed by Inmon & Hackathorn, Widom, Berson, Mattison, Barquin, Berson & Smith, Devlin, Adamson & Venerable, and Turban & Aronson. Some of these definitions follows.

Inmon and Hackathorn (1994) said, "A data warehouse is a subject-oriented, integrated, time-variant, and non volatile collection of data used in support of management's decision making process."

Berson and Smith (1997) asserted that a data warehouse is not a product rather it is an environment. They have defined the DW as a blend of technologies and components the aimed at the integration of operational databases into an environment that permits the strategic use of data.

Devlin (1997) said, "A DW is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context."

Adamson and Venerable (1998) said, "while most computer systems are designed to capture data, data warehouses are designed for getting data out. It is all about getting answers to business questions."

Turban and Aronson (1998) explained that the purpose of the DW is to establish a data repository that prepares the operational database in organizations in an accessible and ready-to-use format for DSS and EIS. Only the data that is required for DSS or EIS is extracted from the operational database and then stored in the DW. Data warehousing or information warehousing as it is sometimes called combines data from different sources into one for end-user access.

DATA WAREHOUSE CHARACTERISTICS

According to (Inmon, 1993) there are four characteristics that generally describe a DW:

1. Time-variant. The DW contains data gathered from different periods. The DW contains a place for storing historical data that can be used for comparisons, trends, or forecasting. Historical data can be up over twenty years old;
2. Non-volatile. The objective of using the DW is to respond to management requests for information. This data is extracted from the operational database and then loaded into the DW database. This means that a data warehouse will always be filled with historical data and should be updated regularly from the operational database. Some DW components are static that is they contain data that does not change over time like a country's past history or events. Whilst another DW components are automatically updated from their sources and they are called active DW component;
3. Subject-oriented. Data are organized according to subject instead of application. Examples of subjects are marketing, production, personnel, sales etc.;
4. Integrated. In many organizations the same piece of data may exist on several databases, to overcome the data redundancy problem there has to be an integration of data sources to avoid duplication.

DW BENEFITS

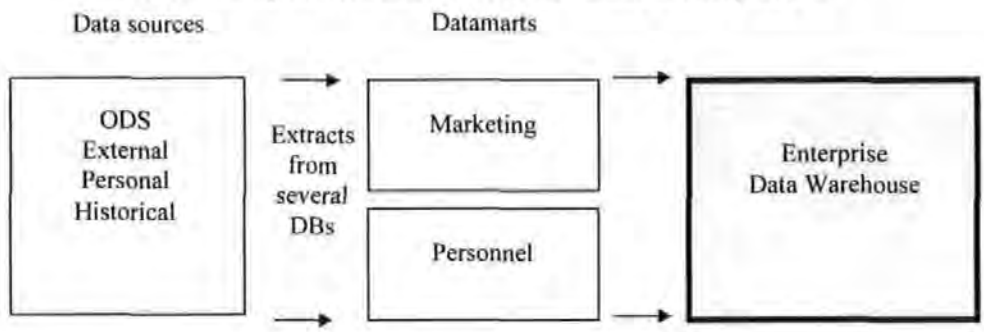
The following tangible benefits have been reported; product inventory turnover is improved, more cost-effective decision making process by separating the query processing from the operational databases, enhancing asset and liability management by providing the overall picture of the enterprise purchasing and inventory transactions, and supporting the corporate strategy that positions the clients at the center of all operations (Cooper et. al, 1999).

DW intangible benefits are (Onder and Nash, 1999); improving productivity by keeping all the required data in a single location, reduces redundant processing and software to enhance DSS applications, enhancing the work process which also affects the success of business process reengineering, improve customer service and organizations will be able to exceed competitor capabilities and achieve competitive advantages.

DATA MARTS

In some cases the DW can contain large number of fields and millions of data records about the entire organization, and in such situations it is called an enterprise DW. A smaller local data warehouse that is classified by subject is called a datamart (Humphries, et. al, 1999; Srivastava and Chen, 1999; Adamson and Venerable, 1998; Hadden, 1998; Edelstein, 1997). Examples of data marts include the marketing data mart and the personnel data mart.

Figure (1)
The relationship between data sources, data marts, and enterprise DW[6]



THE DIFFERENCE BETWEEN THE ODS AND DW

Organizations use TPS to store their business transactions. These TPS have a certain database design requirements and for this reason these types of databases are known as the *ODS or OLTP*. They are optimized for carrying high speed and large numbers of users and transactions. Thousands of users might be connected to the same ODS performing millions of transactions, whilst in the DW there are typically a few users connected to the DW primarily for analysis and complex query purposes (Edelstein, 1997).

Table (1)
ODS Vs DW[7]

Criteria	*ODS*	*DW*
-Purpose	Transaction needs	Strategic needs
-Clients	Users, Administrators	Executives, Managers
-Systems type	Batch	DSS, EIS
-Content	Current values	Summaries, Subsets
-Data actions	Create, Read, Update, Delete, Print	Read-Only
-Data sources	Internal	Internal, Archival, External and Personal
-Size in bytes	Small (MB to GB)	Large (GB to TB)
-Orientation	Application	Strategic
-Response time	Fast (seconds)	Slow (Seconds/Minutes)
-Integration	Partially	Fully

THE STAR SCHEMA STRUCTURE

There is a primary difference between a database that is designed for operational systems (e.g. stock levels system, sales system, and payroll systems) and the database design for the data warehouse. The ODS databases provide the DW with a source of data, however, they lack the functions required to perform efficient analysis and produce reliable results that decision makers really need (Livingston and Rumsby, 1997). The contents of the DW are relatively stable whilst the contents of the ODS change as each transaction initiated.

[6] Adapted from (Adriaans and Zantinge, 1996).
[7] Adapted from (Hadden, 1998).

The best way to build the DW database is by using the star schema structure (sometimes referred to *as multi-dimensional data modelling*-MDDM). The Star schema or MDDM captures the *measurements* of importance to the business and the *parameters* by which the business measurements are broken out. It is a direct reflection on how business processes happen. The measurements are referred to as FACTS, whilst the parameters by which a measurement can be viewed are called DIMENSIONS (Sorensen and Alnor, 1999; Kimball, 1998; Firestone, 1998; Adamson and Venerable, 1998).
A simple star consists of group of tables that describe the dimensions of the business arranged around a central table that contains the business facts. The smaller outer tables are the points of the star, the larger table in the center is the star from which the points radiate. The star schema relies on two major components the *facts* and the *dimensions*. Sometimes for the purpose of enhancing the performance of the DW summary tables are created. Also to transform the ODS database structure to a star schema structure a de-normalization process takes place. Indexing is another technique by which the DW performance can be leveraged. A discussion of these components and techniques follows in the following sections.

Fact tables
This is the central table and generally it is the biggest table in the DW database in terms of records. The DW DB can contain one or more fact tables. When more than one fact table exist, they are called a *fact table family* or sometimes called a *multi-star structure* (Sorensen and Alnor, 1999; Humphries, 1999; Kimball, 1998; Adamson and Venerable, 1998; Livingston and Rumsby, 1997). Examples of fact tables are sales, orders, budgets, shipments, students, and accounts. An important factor in designing the fact table fields is to make them as small as possible in terms of the data size. This is because the size of the fact table will grow dramatically and frequently stores millions of records. Fact tables are built with a multi part primary key, with the key typically consisting of more than one field. Each field points to a matching field in a dimension table. Through these links referential integrity is enforced.

Dimension tables
Dimension tables are the points of the star or the fact table. Each dimension table has a fixed number of records, for example list of courses, list of products, list of employees, or list of the markets. Examples of the dimension tables are time, markets, products, courses, staff members, majors, and vendors. Dimension tables use both character and numeric data types so their fields are much bigger in size than the fact table fields. The number of rows in the dimension table is less than the number of rows in the fact table. Typically, the dimension table contains single-part primary key.

The TIME Dimension
Since the DW includes offloading historical data from the operational systems, so each fact in the DW must be time-stamped (Humphries, et. al, 1999). This requires each DW and/or data mart to include a TIME dimension in the design (Kimball, 1998; Firestone, 1998; Hadden, 1998).

The Granularity of the Fact table
The term GRANULARITY (sometimes referred to as GRAIN of the Fact table) describes the level of detail stored in the fact table. The granularity of the fact table follows the level of detail of its related dimensions (Humphries, et. al, 1999; Kimball, 1998; Adamson and Venerable, 1998; Hadden, 1998). Determining the grain of the fact table is a very critical decision. Granularity at too high level prevents the users of the DW from drilling down into further details of the data. Granularity at a low level of detail results in an enormous increases in the DW size and consequently affects both cost and performance. *For example*, if the record in the Time dimension table represents a semester, the record in the Colleges dimension table represents a College, the record in the Majors dimension table represents a Major, and the record in the Students_GPA fact table represents a student then the grain of the fact table in connection to these dimensions would be: the student GPA per College per Major per Semester.

Summary tables
A summary table is a DW table that includes data frequently retrieved by users. Instead of searching in the entire fact table a snapshot is taken and stored in a summary table, when the user is asking the query the result comes from the summary table (Hadden, 1998). A summary table allows the DW to respond rapidly to known or anticipated business queries. A survey (1999) was done by Compaq Corporation to examine MS-SQL Server data marts tools. Results showed that it is not recommended to create summary tables if the query is used infrequently or if the data retrieved by this query is more than 20% of the rows in the fact table. In these cases it is more efficient to retrieve these data from the fact table directly.

De-normalization

Unlike the relational structure, the star schema structure uses the *de-normalization* process to enhance the performance of the DB tables. De-normalization helps to reduce the number of joins between the tables, thus making the query writing process easier, and also to speed up the query (Hadden, 1998; Sorensen and Alnor, 1999). Whereas the normalization process tries to split-up tables, the de-normalization process rolls-up all the data about the dimension in one table.

Indexing

Since the DW is built to support the strategic use of data the DSS/ EIS will be extracting information often using queries that require extensive amounts of processing time. In order to reduce the processing time the DW designer makes extensive use of indexing. Users expect fast response, but as data and usage expand the DW will not able to respond quickly to all users' queries. An index functions like a smaller table in ordered sequence (Paller, 1997). To ensure an optimal indexing methodology, multiple indexes are created on most of the dimension tables. Another strategy to ensure fast access to the rows of the dimension tables is to put an index on each field (Livingston and Rumsby, 1997; Berson and Smith; 1997). The most relevant indexing approach for the DW is to index all columns in the dimension tables and all foreign keys in the fact table (Kimball, 1998; Livingston and Rumsby, 1997) which will result in performance improvements and enhance query processing speed.

DATA WAREHOUSE COMPONENTS

Data source

These are the ODS' databases, external, personal or archival data sources. Different data source formats can be used as sources of data, examples like VSAM, IMS, RMS, DB2, relational, flat files or other formats (Mimno, 1997);

Data extraction and transformation tools

These are used to extract data from the data source files and clean up the data then be sure that all the relevant data required by the users is available in the DW. The extraction can be done using a standard RDBMS (ORACLE, SYBASE, INFORMIX, DB2, SQL, ACCESS). The data transformations might include aggregating, inserting default values, sampling or summarizing the data to reduce the size of the DW (Widom, 1995). Various tools eg. Extract, Integrity Data Reengineering, Platinum Warehouse have been developed to support these transformation processes. During the transformation process the operational system fields/attributes are copied to the DW. Many transformation types are available to perform this mapping; field splitting, field consolidation, standardization, and de-duplication.

Data modeling tools

These tools are used to prepare the DW structure from both the data source and the target data warehouse database;

Central repository

This component is used to store the metadata (data about data). Metadata describes the transformation between the source and the target databases.

Target DB

This is the DW database, where the data of interest will be stored. The target database can be a conventional relational database, proprietary, or multi-dimensional (Mimno, 1997). In many cases organizations use the standard RDBMS to build the DW target DB.

Front end

These are the tools used to analyze the data stored in the DW database. These tools include; General-purpose relational data access, Data mining tools, DSS, EIS, and Web tools that perform search and query in the WWW environment. For some applications a mix of the front-end tools may be required (Laudon and Laudon, 2000; Mimno, 1999; Mattison, 1997; Berson and Smith, 1997; Mimno, 1997; Adriaans and Zantinge, 1996).

CLIENT/SERVER STRUCTURES FOR SUPPORTING DATA WAREHOUSING

Client/Server architecture consists of a number of workstations (Clients), one or more higher configuration workstation (Server(s)), and a local area network (LAN) connecting them all together (Edwards, 1999; Orfali, 1999; Delis and Roussopoulos, 1992). Of all the techniques currently available, client/server represents the best

choice for building a data warehouse (Orfali, et al., 1999; Edelstein, 1997; Berson and Smith, 1997; Delis and Roussopoulos, 1992). The role of the client depends on the DW architecture. There are two basic architectures.

The two-tier DW

The two-tier (*2-tier*) DW architecture or as sometimes called *the fat client* model (Edwards, 1999; Edelstein, 1997), in which clients' functions include GUI presentation logic, query definition, data analysis, report formatting, summarization, and data access, while the DW server performs data logic, data services, metadata maintenance and the file services. The two-tiered architecture lacks scalability and flexibility. As the number of users increases the data access requirements imposes heavy burden on the server and the performance degrades (Mimno, 1997). Source data and the data warehouse DB reside on the server (tier 2), whilst business rules that are shared across the organization and graphically oriented end-users run on LAN-based workstations (tier 1).

The multi-tier DW

The multi-tier (*3-tier*) DW architecture or as sometimes called *the thin client* model, handles the scalability and flexibility problems through the application servers. Application servers perform data filtering, summarization, aggregation, support metadata, data access, and provides multi-dimensional views. The multi-tier architecture reflects the multi-tier client/server model (Edwards, 1999; Edelstein, 1997; Dewire, 1998). Source data resides on the server (tier 3), data warehouse DB and business rules that are shared across the organization are stored in a DB Server (tier 2), and graphically oriented end-users run on LAN-based workstations (tier 1).

When a client/server was a departmental or campus-based phenomenon, the shortcomings of 2-tier were not very important. They certainly did not outweigh the advantages provided by 2-tier's ease of development. But as client/server grew up to run mission critical applications-especially those of intergalactic proportions- 3-tier became essential" Edwards, 1999

THE DATA WAREHOUSE DEVELOPMENT APPROACHES

Two basic approaches are used to build a data warehouse approaches (Edelstein, 1997; Berson, 1996).

1. *The 'top down' approach.* The organization has developed an enterprise data model, collected enterprise-wide business requirements. It then builds an enterprise data warehouse with subset data marts.
2. *The 'bottom up' approach.* This implies that the business priorities result in developing individual data marts, which are then integrated into the enterprise data warehouse.

The bottom up approach is probably the more realistic, but the complexity of the integration may become a serious obstacle, and the warehouse designers should analyze carefully each data mart for integration purpose. If the set up of a datamart is to be used with data mining techniques, then optimizing the local databases is important.

USERS OF THE DATA WAREHOUSE

Taking into consideration the nature and characteristics of the DW and its strategic use and the historic data it contains reveals that the typical users are; those who need certain amount of data in a special format for the reason of summarizing, and aggregating data, those who deal with analyzing and displaying historical data, those who need to reply to the frequently asked queries-FAQ, and those who need continuous accessing of a certain data and exception reporting (Taha, et al., 1997). This list reveals the fact that *managers* and *executives* (decision makers in organizations) are the primary candidates to use the DW (Berson and Smith, 1997; Adriaans and Zantinge, 1996).

THE SURVEY

The population of this questionnaire is the Egyptian universities. These universities are classified into two groups government and private-funded. According to the Egyptian Supreme Council of Universities statistics (1999), The UNESCO World List of Universities (2000), and The British Council Global Education and Training Information Service-Egyptian Universities (1999) there are twenty-one Egyptian universities; eight are private and thirteen are government. Only 13 universities agreed to participate in the survey, and then a number of 670 questionnaires were sent to 13 universities (See Appendix A). The Admission and Registration Decision Support Systems questionnaire (ARDSS) is the instrument used for data collection. The ARDSS overall Reliability Alpha coefficient is 0.96.

Response rate= 167/670= 24.9 %. This response rat is adequate in this kind of research that is based on mailed questionnaires (Chen, et. al, 1998; Saunders, et. al, 1997; Teo and King, 1996). The position contribution to the response rate and the university type contribution is illustrated in the following tables.

Table (2)
The respondents

Position	Responses	Per %
Dean	27	16.17
Associative Dean	30	17.96
Registrar	48	28.74
Admission Officer	10	5.99
Others; Director, Senior Academic Advisors	52	31.14
Total	167	100 %

Table (3)
The respondents

University	Responses	Per %
Government	106	63.5
Private	61	36.5
Total	167	100 %

The ARDSS survey instrument is a multi-part questionnaire. The ARDSS instrument consists of 85 questions in two sections; firstly describing current Admission and Registration Information Systems running in the Egyptian Universities, and secondly deriving the information needs for a new system from which the DW is expected to be part of.

From the ARDSS instrument 85 questions, only 8 questions are about the existence and the use of the DW within the universities. Examples of these questions are found in Appendix (B). Results show that there is no single university in Egypt has a DW. However, results also indicate that there is a need for the DW technology within the Admission and Registration functions in universities. 86.2 % of respondents think that their ideal Admission and Registration Information Systems should be linked to a DW. Based on these needs derived from the ARDSS questionnaire, the following DW was developed.

DEVELOPMENT OF THE DW FOR THE EGYPTIAN UNIVERSITIES

The Star Schema Structure

Refer to Appendix (C) for details. The data warehouse for the Admission and Registration functions in universities depends on the operational system. The next matrices show how the operational system files has been mapped into the DW files.

Table (4)
Source to Target Field Matrix

	NO	1										
	SCHEMA											
	TABLE	FACT: EXAM_RESULT										
TABLE	FIELD											
EXAM_ RESULT	EXA_RES_COUNTER PK[8] EXA_RES_DATE EXA_RES_MARK	X	X	X								
	NO	2										
	SCHEMA											
	TABLE	FACT: STUDENT_RECORD										
TABLE	FIELD											
STUDENT_ RECORD	STU_REG_CODE PK /STU_REG_GPA /STU_REG_TOTAL_HOURS /STU_REG_TOTAL_ACHIEVED STU_REG_STATUS STU_REG_PENALTY STU_REG_DATE	X	X	X	X	-	-	X				
	NO	3										
	SCHEMA											
	TABLE	DIMENSION: COURSE_BOOKING										
TABLE	FIELD											...

[8] PK stands for the primary key of the entity in the OLTP system, but not the PK of either the FACT or DIMENSION table.

COURSE_ BOOKING	BOO_SERIALNO PK	X										
	BOO_DATE		X									
	/BOO_TOTAL_COURSES			X								
	/BOO_TOTAL_HOURS				X							
	NO	4										
	SCHEMA											
	TABLE	DIMENSION: APPLICANTS										
TABLE	FIELD											
APPLICANTS	APP_NAME	X										
	APP_GENDER		X									
	APP_TELEPHONE			-								
	APP_PERCENTAGE				X							
	APP_ADDRESS					-						
	APP_BIRTH_DATE						X					
	APP_CODE PK							X				
	APP_SEC_CERT_YEAR								X			
NATIONALITY	NAT_NUMBER PK									X		
	NAT_NATIONALITY										X	
SECONDARY_ CERTIFICATES	SEC_CODE PK											X
	SEC_NAME											X
	SEC_TYPE											X
	SEC_ORIGIN											X
BATCH_ DATA	BAT_CEILING											X
	BAT_CODE PK											X
	BAT_YEAR											X
	BAT_SEMESTER											X
	NO	5										
	SCHEMA											
	TABLE	DIMENSION: COURSES										
TABLE	FIELD											
COURSE	COU_TITLE	X										
	COU_CODE PK		X									
	COU_CREDIT_HOURS			X								
	COU_PASS_MARK				X							
	COU_FULL_MARK					X						
	COU_STATUS						X					
	COU_PREREQUISITE							X				
	COU_LABS								X			
	COU_STAGE									X		
	COU_AREA										X	
COURSE_ GROUPS	GRO_CEILING											X
	GRO_LETTER											X
	GRO_CODE PK											X
PREREQUISITE	PRE_SERIAL PK											X
	PRE_DETAILS											X
	PRE_COURSE_1											X
	PRE_COURSE_2											-
	PRE_COURSE_3											-
	PRE_COURSE_4											-
	PRE_COURSE_5											-
COURSE_ MAJOR	COU_MAJ_COUNTER PK											X
	NO	6										
	SCHEMA											
	TABLE	DIMENSION: FEES										
TABLE	FIELD											

[9] Due to space constraint, this column might be used to represent more than one field.

TUITION	TUI_COUNTER PK	X										
	TUI_AMOUNT		X									
	TUI_CURRENCY			X								
	TUI_TERM				X							
SEMESTER	SEM_CODE PK					X						
	SEM_TITLE						X					
	SEM_YEAR							X				
	NO	7										
	SCHEMA											
	TABLE	DIMENSION: EXAMS										
TABLE	FIELD											
EXAM	EXA_CODE PK	X										
	EXA_TITLE		X									
	EXA_TYPE			X								
	EXA_SHELTER				X							
	NO	8										
	SCHEMA											
	TABLE	DIMENSION: STUDENTS										
TABLE	FIELD											
STUDENT	STU_REGISTRATION_NO PK	X										
	STU_STATUS		X									
AUTHORITY	AUT_NAME			X								
	AUT_ADDRESS				-							
	AUT_TELEPHONE					-						
	AUT_CONTACT_PERSON						-					
	AUT_CODE PK							X				
STUDENT_ ASSISTANTSHIP	STU_ASS_DATE											X
	STU_ASS_COUNTER PK											X
	NO	9										
	SCHEMA											
	TABLE	DIMENSION: PAYMENTS										
TABLE	FIELD											
PAYMENT_ RECORD	PAY_AMOUNT	X										
	PAY_CURRENCY		X									
	PAY_METHOD			X								
	PAY_RECEIPT_NO PK				X							
	NO	10										
	SCHEMA											
	TABLE	DIMENSION: ASSISTANTSHIPS										
TABLE	FIELD											
ASSISTANTSHIP_ TYPE	ASS_TITLE	X										
	ASS_SERIAL PK		X									
	ASS_CATEGORY			X								
	ASS_REQUIREMENTS				X							
	ASS_ADVANTAGES					X						
	ASS_DISCOUNT						X					
	NO	11										
	SCHEMA											
	TABLE	DIMENSION: COLLEGES										
TABLE	FIELD											
COLLEGE	COL_NAME	X										
	COL_CODE PK		X									
	COL_LOCATION			X								
DEPARTMENT	DEP_TITLE				X							
	DEP_ID PK					X						
	DEP_LOCATION						X					
MAJOR	MAJ_TITLE							X				
	MAJ_SNO PK								X			

	MAJ_MIN_SEC_PERCENT									X		
	MAJ_DETAILS										X	
	NO	12										
	SCHEMA											
	TABLE	DIMENSION: STAFF										
TABLE	FIELD											
STAFF	STA_NAME	X										
	STA_ID PK		X									
	STA_ADDRESS			-								
	STA_HIRING_DATE				X							
	STA_TITLE					X						
	STA_POSITION						X					
	STA_SOCIAL_STATUS							X				
	STA_GENDER								X			
	STA_CONCENTRATION									X		
	STA_QUALIFICATIONS										X	
	STA_EXPERIENCE											X
	NO	13										
	SCHEMA											
	TABLE	DIMENSION: TIME										
TABLE	FIELD											
	DATE PK	X										
	DAY_OF_WEEK		X									
	MONTH			X								
	QUARTER				X							
	YEAR					X						

Example SQL statements that could be used to derive information from the DW:

-Student_Applicant _Booking:

```
SELECT applicant_student.stu_registration_no, applicant_student.app_name,
BOOKING_DIMENSION.boo_cou_code, BOOKING_DIMENSION.boo_date
FROM applicant_student INNER JOIN BOOKING_DIMENSION ON
applicant_student.stu_registration_no = BOOKING_DIMENSION.boo_stu_reg_no
GROUP BY applicant_student.stu_registration_no, applicant_student.app_name,
BOOKING_DIMENSION.boo_cou_code, BOOKING_DIMENSION.boo_date
WITH OWNERACCESS OPTION;
```

-GPA:

```
SELECT exam_result_fact_gpa.stu_key,
[exam_result_fact_gpa]![SumOfexa_mark]/[exam_result_fact_gpa]![CountOfcou_key] AS Expr1
FROM exam_result_fact_gpa
GROUP BY exam_result_fact_gpa.stu_key,
[exam_result_fact_gpa]![SumOfexa_mark]/[exam_result_fact_gpa]![CountOfcou_key]
WITH OWNERACCESS OPTION;
```

-Exam_Result_Application:

```
SELECT EXAM_DIMENSION.exa_title, EXAM_DIMENSION.exa_type,
EXAM_DIMENSION.exa_key, EXAM_RESULT_FACT.exa_mark,
APPLICANT_DIMENSION.app_name
FROM EXAM_DIMENSION INNER JOIN (APPLICANT_DIMENSION INNER JOIN
EXAM_RESULT_FACT ON APPLICANT_DIMENSION.app_key = EXAM_RESULT_FACT.app_key)
ON EXAM_DIMENSION.exa_key = EXAM_RESULT_FACT.exa_key
WITH OWNERACCESS OPTION;
```

CONCLUSION

-The DW technology is a valid option and is expected to enhance the managers' decision quality when employed to the Admission and Registration Information Systems in Universites.

-It is therefore recommended to be utilized by universities.

REFERENCES

Adamson, C., and Michael Venerable. (1998). Data Warehouse Design Solutions, WILEY.

Adriaans, P., and Dolf Zantinge. (1996). Data Mining, Addison Wesley Longman.
Arab Academy for Science and Technology & Maritime Transport (AASTMT) Statistics. Different volumes from 1990-1997.
Arab Academy for Science and Technology & Maritime Transport. Admission and Registration records, October 1995.
Barquin, R. (1997). A Data Warehousing Manifesto. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.3-16.
Barquin, R., Paller, A., and Herb Edelstein (1997). Ten mistakes to avoid for data warehousing managers. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.145-156.
Berson, A. (1996). Client/Server Architecture, McGraw-Hill.
Berson, A., and Stephen Smith. (1997). Data Warehousing, Data Mining, & OLAP, McGraw-Hill.
Cooper, B., Watson, H., Wixom, B., and Dale L. Goodhue. (2000). Data Warehousing Supports Corporate Strategy At First American Corporation. MIS Quarterly, vol. (24:4), Dec. 2000, pp.547-567.
Date, C. (1995). An Introduction to Database Systems, Addison-Wesley.
Delis, A., and Nick Roussopoulos. (1992). Performance and Scalability of Client-Server Database Architecture. In Proceedings of the 18th VLDB Conference Vancouver, British Columbia, Canada.
Devlin, B. (1997). Data Warehouse: From Architecture to Implementation, Addison-Wesley.
Dewire, D. (1998). Thin Clients, McGraw-Hill.
Edelstein, H. (1997). An introduction to data warehousing. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.31-50.
Edwards, J. (1999). 3-Tier Client/Server At Work, Wiley.
Elmasri, R., and Shamkant B. Navathe. (1994). Fundamentals of Database Systems, The Benjamin/Cummings Publishing.
Fayyad, U., Piatetsky, G., and Padharic Smyth, (1996). From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.1-34.
Firestone, J. (1998). Dimensional Modeling and E-R Modeling in the Data Warehouse, EIS Inc.
Funkhouser, T. (1995). RING: A Client-Server System for Multi-User Virtual Environment. In Proceedings of the ACM SIGGRAPH Conference New York, USA, pp.85-92.
Garcia-Molina, H., Labio, W., and Jun Yang. (1998). Expiring Data in a Warehouse. In the Proceedings of the 24th VLDB Conference, New York, USA, 1998.
Global Education and Training Information Service-Egyptian Universities, The British Council, 1999.
Hackett, G. and Peter Luffrum. (1999). Business Decision Analysis, Blackwell Publishers.
Hadden, E. (1998). Building for Successful Data Warehouses and Data Marts, BBS International.
Han, J., Fu, Y., Wang, W., Krzysztof, K., and Zaiane, O. (1996).DMQL: A data mining query language for relational databases. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.
Han, J., Lakshmanan, L., and Raymond T. Ng. (1999). Constraint-based, multidimensional Data mining. IEEE Computer, Aug 1999, pp.46-50.
Hufford, D. (1997). Metadata Repositories: The key to unlocking information in Data Warehouses. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.225-262.
Humphries, M., Hawkins, M., and Michelle C. Dy. (1999). Data Warehousing Architecture and Implementation, Prentice Hall.
IEF, CASE tools manuals, by Texas Instruments. Volume 1995.
Inmon, W.(1993). Building the Data Warehouse. QED publishing Group.
Inmon, W., and Richard D. Hackathorn. (1994). Using The Data Warehouse, John Wiley & Sons.
Jordan, D., and P. Smith. (1997). Mathematical Techniques, second edition, Oxford.
Kimball, R. (1998). The Data Warehouse Toolkit, John Wiley and Sons.
Klein, M., and Leif B. Methlie. (1995). Knowledge-Based Decision Support Systems, John Wiley & Sons.
Lee, A., and Elke A. Rundensteiner. (1998). Data Warehouse Evolution: Consistent Metadata Management. In proceedings of the IEEE Conference on SMC, San Diego, California, USA.
Leidner, D., and Mark Fuller. (1997). Improving student learning of conceptual information: GSS supported collaborative learning vs. individual constructive learning. Decision Support Systems, vol. (20), pp.149-163.
Livingston, G., and Bob Rumsby, (1997). Database Design for The Data Warehouses: The Basic Requirements. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.179-198.

Long, L. (1989). Management Information Systems, Prentice Hall.
Mattison, R. (1997). Data warehousing and data mining for telecommunications, Artech House.
McFarland, G., Rudmik, A., and Modus Operandi. (1999). Object-Oriented Database Management Systems Revisited, DACS.
Mi, P., and Walt Scacchi. (1996). A meta-model for formulating knowledge-based models of software development. Decision Support Systems, vol. (17), pp.313-330.
Mimno, P. (1997). Data Warehousing Architectures. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.159-177.
Mimno, P. (1999). Build your Data Warehouse Right the First Time, Brio Technology. URL: http://www.brio.com
Neal, D. (1997). How to Justify the Data Warehouse and Gain the Top Management Support. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.91-115.
O'brien, J. (1996). Introduction To Information Systems, IRWIN.
O'Driscoll, T., Massey, A., and Mitzi M. Montoya-Weiss. (1999). Virtual Mentor: Enabling Knowledge Management Through An Electronic Performance Support System, SIM International paper award competition-2nd place. URL: http://www.simnet.org
Onder, J., and Todd Nash. (1999). The Approach to Building a Business Data Warehouse, SYSIX. URL: http://www.sysix.com
Orfali, R., Karkey, D., and Jeri Edwards. (1999). Client/Server Survival Guide, John Wiley and Sons.
Paller, A. (1997). A Roadmap To Data Warehousing. In Barquin, R., and Herb Edlestein. Planning and Designing The Data Warehouse, Prentice Hall, pp.17-29.
Parsaye, K., Chignell, M., Khoshafian, S., and Harry Wong. (1989). Intelligent Databases: Object-Oriented, Deductive, Hypermedia Technologies, WILEY.
Pyle, D. (1999). Data preparation for Data Mining, Morgan Kaufmann Publishers.
Saraee, M., and B. Theodoulidis, (1995). Knowledge Discovery in Temporal Databases. In Colloquium on Knowledge Discovery in Databases, IEE, UK, pp. 1/1-1/4.
Silberschatz, A., and Alexander Tuzhilin. (1995). On Objective Measures of Interestingness in Knowledge Discovery. In Proceedings of KDD Annual Conference, Montreal, Canada.
Silberschatz, A., and Alexander Tuzhilin. (1996). User Assisted Knowledge Discovery: how much should the user be involved. In Proceedings of SIGMOD Annual Conference, The University of Columbia, USA.
Sorensen, J., and Karl Alnor. (1999). Creating Data Warehouse Using SQL Server. In the proceedings of the International Workshop on Design and Management of Data Warehouses, DMDW'99.
Srivastava, J., and Ping-Yao Chen. (1999). Warehouse Creation-A Potential Roadblock to Data Warehousing. IEEE Transactions on Knowledge and Data Engineering, vol. (11:1), pp.118-126.
Staudt, M., Vaduva, A., and Thomas Vetterli. (2000). The Role of Metadata for Data Warehousing, Swiss Federal Office of Professional Education and Technology.
Taha, Y., Helal, A., and Ahmed, K. (1997). Data Warehousing: Usage, Architecture, and Research Issues. ISMM Microcomputer Application Journal, vol. (16:2), pp.1-8.
Teklitz, F., Krneta, P., and Russ Puryear. (1999). SYBASE Business Intelligence on the Web with Windows NT, SYBASE BI. URL: http://www.sybase.com/bi
The Egyptian Supreme Council of Universities Statistics, The Egyptian Supreme Council of Universities, 1999.
Travis, D. (1998). Thin Clients: Web-Based Client/Server Architecture and Applications, McGraw-Hill.
Turban, E. (1993). Decision Support and Expert Systems, Management Support Systems, Macmillan.
Turban, E., and Jaye Aronson, (1998). Decision Support Systems and Intelligent Systems, Prentice Hall.
Whitten, J., Lonnie D. Bentley, and Victor M. Barlow. (1994). Systems Analysis And Design Methods, IRWIN.
Widom, J. (1995). Research Problems in Data Warehousing. In the Proceedings of 4th International Conference on Information and Knowledge Management (CIKM), pp.25.30.
World List of Universities and other Institutions of Higher Education, MACMILLAN, twenty second edition, 2000.
Zhang, X., and Elke A. Rundensteiner. (1998). Data Warehouse Maintenance Under Concurrent Schema and Data Updates. Computer Science Technical Report Series, Aug 1998, pp.1-29.

APPENDIX (A)

Response rate

	Dean		Asso. Dean		Registrar		Adm. Officer		Others; Sen. Acd. Advisor or Director		Total
University-level						5		7		35	47
Faculty of Science	4		5		3				3		15
Faculty of Commerce	2		2		5				1		10
Faculty of Law	3		1		4						8
Faculty of Hotels and Tourism	1		1		1						3
Faculty of Education	2		7		6						15
Faculty of Medicine	1		1		2						4
Faculty of Physical Education					2						2
Faculty of Dentist			1		1						2
Faculty of Pharmacy	2		2		1				1		6
Faculty of Veterinary	1				1						2
Faculty of Arts	1		2		2				3		8
Faculty of Agriculture	5		3		3				1		12
Faculty of Engineering	2		2		2	1		1		1	9
Faculty of Home Economics	1		1						1		3
College of Management				1		6		2			9
Faculty of Informatics	1										1
Faculty of Social Services	1		1		1						3
College of Marine Engineering						1					1
College of Maritime Studies						1					1
DSS Unit									6		6
Total respondents	27		29	1	34	16		10	16	36	167

APPENDIX (B)

The ARDSS Instrument Sample Questions (Questions are all scaled to either Yes or No)

Q.5 *Is your information system linked to an archival or historical students' database?*

....

Q.7 *Do you encounter situations where your decision will be enhanced if you search in the students' history before making the decision?*

...

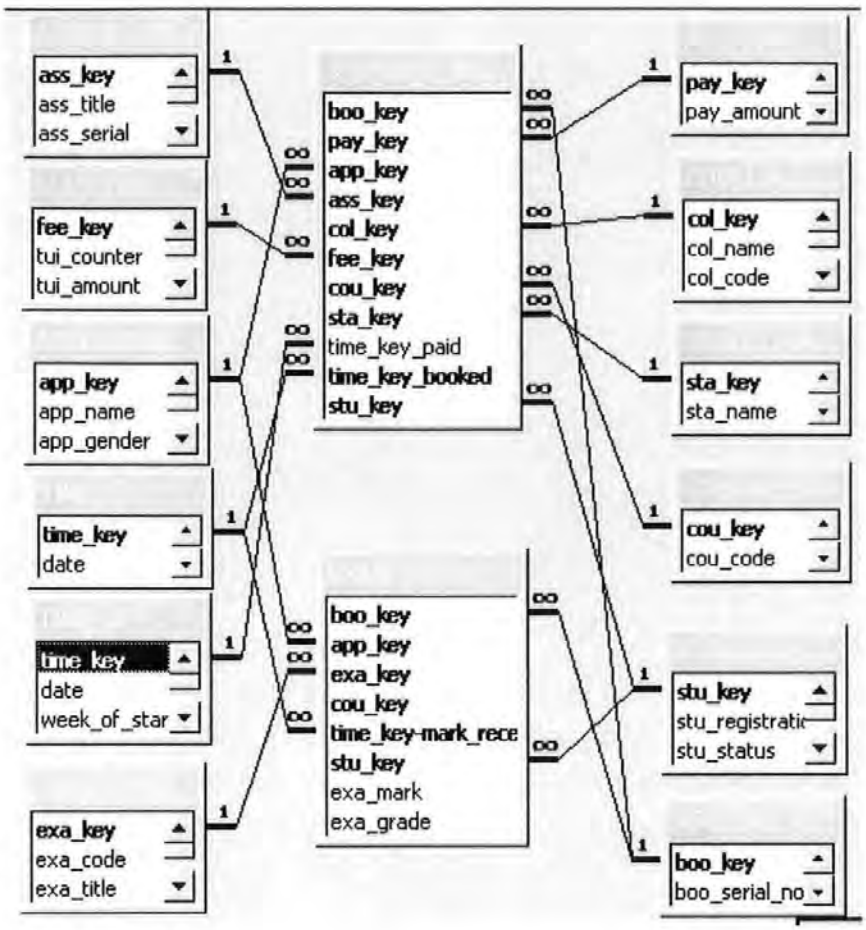
Q.24 *Given the following list of admission and registration functions, please choose Yes for those functions that should be part of the ideal admission and registration information system, or No for those that should not be part of it. Use the sign (✓) to make your choice. Remember that there are no right or wrong answers:*

....

F. Using historical data

....

APPENDIX (C)
The DW Star Schema Structure



Developing Star Schema Structures- A Practical Study Applied To Egyptian Universities[10]

Ahmed El-Ragal Terry Mangles Ian Chaston

Business Management department

University of Plymouth Business School

Drake Circus, Plymouth, PL4 8AA

Devon, UK

Fax: +44-01752-232853

Tel: +44-01752-232850 Tel: +44-01752-232852 Tel: +44-01752-232810

a.el-ragal@plymouth.ac.uk terry.mangles@pbs.plym.ac.uk ian.chaston@pbs.plym.ac.uk

Abstract

This paper investigates many of the practical issues surrounding the development and implementation of a star schema structured data warehouse. The paper introduces and summarizes the organizational requirements that are required to underpin the student recruitment process in higher education. These requirements have been identified following an in-depth survey of the recruitment process in Egyptian Universities. This survey was used to identify the required data sources together with the likely users and their information needs. The survey was sent to senior managers within the Egyptian Universities (both private and public) with responsibility for student recruitment, in particular the admission and registration processes. Further, access to a large database has allowed us to test the practical suitability of using a data warehouse structure and knowledge management tools within the decision-making framework. The design of the proposed data warehouse will be developed and illustrated using CASE tools. It is not the intention of this paper to compare tools but to illustrate the use and benefits that these can accrue to the systems developer. In particular, the benefits of matrix verification within the data warehouse design process will be explored and the paper will illustrate how these tools can be used to produce efficient and bug free data warehouse. Finally, the paper will discuss the use of front-end tools to develop an easy to use system without which the data warehouse would not be used.

[10] The BIT World 2001 Conference, in Cairo, Egypt, 4-6 June 2001, pp. 128.

Developing a new Decision Support System for University Student Recruitment[11]

Authors: **Terry Mangles**
Principal Lecturer in Information Management
Ahmed El-Ragal
Lecturer in Information Systems

Terry Mangles
University of Plymouth Business School
University of Plymouth
Drake Circus
Plymouth, PL4 8AA
Devon, UK

Tel: +44-01752-232852
Fax: +44-01752-232853
E-mail:terry.mangles@pbs.plym.ac.uk

Ahmed El-Ragal
Arab Academy For Science and Technology
Management Information Systems Department
P.O. Box 1029, AASTMT, Miami
Alexandria, EGYPT

Tel: +2-03-5485473
Fax: +2-03-5566072
E-mail:aelragal@mis.aast.edu

Track:
Decision Support System

Extended abstract

Universities are facing increasing pressures in order to meet student expectations, educational ratings, and financial efficiency gains. Student expectations have been heightened by the fact that they are contributing to the cost of their education and the amount of information now readily available about Universities e.g. location, social life, degree classification, accommodation etc. As far as educational ratings are concerned the UK government now produces league tables of universities in term of their teaching and research records whilst in Egypt a concept of grade point average is used to determine 'value added'. Financial pressures have increased as efficiency gains have been implemented resulting in the need to ensure that student progression rates are improved or alternatively, student recruitment is increased in order to maintain cash flow. In an increasingly competitive market place increasing student recruitment to support high failure rates is proving to be unsustainable and therefore methods are being sought to improve progression rates. In fact, improving the student progression rate could have three immediate benefits:

- Improve the financial position of the University particularly as far as planning processes are concerned;
- Improve the perceived quality of the University courses by increasing the success rate;

[11] To appear in proceedings of the ICEB Conference, in Hong Kong, 19-21 December 2001.

- Reduce pressures on the recruitment process.

This paper investigates how developing a decision support system could lead to more informed and hence better decisions regarding recruitment can be mad to ensure retention and successful completion of their chosen course of study.
The paper introduces and summarizes the organisational requirements that are required to underpin the student recruitment process in higher education. These requirements have been identified following an in-depth survey of the recruitment process in Egyptian Universities. This survey was used to identify the required data sources together with the likely users and their information needs. The survey was sent to senior managers within the Egyptian Universities (both private and public) with responsibility for student recruitment, in particular the admission and registration processes. Further, access to a large database has allowed the testing of the practical suitability of using a data warehouse structure and knowledge management tools within the decision making framework.

The paper uses this practical analysis together with the definition of authors such as Sorensen & Alnor (1999); Humphries (1999); Adamson & Venerable (1998); Mattison (1997); Barquin (1997); Berson & Smith (1997); Devlin (1997); Berson (1996); Kimball (1996); Widom(1995); and Inmon & Hackathorn (1994) to develop a data warehouse definition . Further the paper will analyze the key characteristics of data warehouse and explore the advantages and disadvantages of such data structures.

The paper then evaluates the work of authors such as Gray and Watson (1999); Cooper et al. (2000) to develop a proposal for the components of a decision support system.

The potential benefits of the data warehouse within the student recruitment process will be examined and the main users of the proposed system identified (Onder & Nash, 1999; Turban & Aronson, 1998). Again this will draw heavily on the results of our survey and in particular will focus on the requirements of senior managers and how a data warehouse can be used to support a decision support system to assist them in making key decisions.

A number of data mining techniques will be explored (Chen et al., 2001; Guyon et al., 1996) and applied to analyzing the data warehouse and how these techniques can be used to discover knowledge.

Finally the paper will review the extent to which the proposed decision support system can be used to improve the effectiveness of the student recruitment process.