2022-01-01

A computational method to track the evolution of business models in the Digital Economy

Wood, Z

Proceedings of the Annual Hawaii International Conference on System Sciences

A computational method to track the evolution of business models in the Digital Economy

Zena Wood
University of Exeter
Z.M.Wood2@exeter.ac.uk

David Walker
University of Plymouth
david.walker@plymouth.ac.uk

Glenn Parry
University of Surrey
g.parry@surrey.ac.uk

Abstract

Companies within the Digital Economy are evolving their business models as they take advantage of the opportunities afforded by emerging digital technologies. There is a need to develop methods that will allow researchers and policy makers to understand the existence of, and relationships between, the different business models within the Digital Economy and track their evolution. Such methods could also help quantify the size and growth of the Digital Economy. This paper presents a computational method, which utilizes machine learning and web scraping, to identify new business models, and a taxonomy of organisations, through the analysis of a firm's webpage. The work seeks to provide an autonomous tool that provides regular output tracking trends in the number of firms in a market, their business model and changes in activity from product to service over time. This information would provide valuable and actionable insight for researchers, firms and markets.

1. Introduction

Companies are evolving their business models, creating new value propositions that reflect the changing needs of customers and utilising the increasing opportunities that are offered by new technologies in the Digital Economy [1]. Standard Industrial Classification (SIC) codes were developed in 1949 as a way of measuring and classifying the economic activity of every business within the county [2]. However, these SIC codes are now out-of-date with one in ten companies in the UK being classified as 'other' [2]. Frameworks have been produced that highlight the key concepts that can describe and differentiate business models (e.g., [3, 4, 5]). However, there is a need to develop methods that will allow researchers and policy makers to understand the existence of, and the relationships between, the different business models within the Digital Economy and track their evolution. Such methods could also help quantify the size and growth of the Digital Economy [2].

The term servitization refers to the change in business models observed through the creation of value by developing additional services to support a firm's offering [6, 7, 8]. The result of the process of servitization leads to a continuum of firm offers from product firms, through product-service systems to high-value knowledge intensive offerings [9]. The ability to measure the extent of servitization over time, and categorise according to the degree of product/service offer, allows changes in service diffusion and firm strategy over time to be analysed, providing valuable insight for firms and markets. Empirical evidence that explores servitization (e.g., [10, 11, 12, 13]) often relies on methodologies that are time consuming (e.g. interviews and manual coding) and on databases which are not controlled by the individual firms of interest so may contain inaccuracies.

In this paper a computational method is presented that seeks to provide a regular output that may be of interest to the research community and organisations, tracking trends in the number of firms in a market, their activity and changes in business model including activity from product to service over time (i.e., servitization). The method incorporates web scraping and machine learning to identify different business models. A basic framework for a business model is adopted which considers three elements: value proposition, realisation of value in use, and worth capture [5]. The method employs computational methods that can be run at minimal time and cost to the research team.

From the analysis of the output of the method a taxonomy of organisations emerges. Currently the taxonomy emerges from analysing the output of the web-scraping component. Going forward the generation of the taxonomy will be fully automated through the use of ontological theory. The flexibility of the method means that it can be applied to any dataset with minimal effort from those that are running it. The ability to

URI: https://hdl.handle.net/10125/79492
978-0-9981331-5-7

repeatedly run the method will provide a regular output tracking trends in the number of firms in a market, their activity and changes in activity from product to service over time. This information would provide valuable and actionable insight for researchers, firms and markets (e.g., help an organisation to strategically determine where to position themselves in a competitive landscape and identify any encroachment, a new strategic insight approach to monitor markets).

The proposed method can be considered more inclusive than existing options since it does not rely on databases where firms must meet a qualification to appear. The data that the method relies upon is created and disseminated by the firm themselves who have a direct commercial interest in its timeliness and accuracy. Data publicly available on web-pages provide a rich source of data that can be useful in identifying trends.

Within this paper, the potential of the proposed method is explored with the use of a case study to show how a set of businesses could be clustered based on their business models and the way in which they describe themselves. The method has been implemented and successfully applied to a sample set of websites: the West Country's 150 biggest businesses [14]. The case study forms a proof of concept and highlights how the method could be improved and evolved. Future work will include advancing the method to use automated identification of keywords and clustering, rigorous testing and its application to a broad range of datasets.

An overview of existing work is given in section 2 followed by a description of the method (section 3). The application of the method to a sample set of webpages is presented in section 4 and the results discussed in section 5. The paper concludes with a discussion of how the method could be extended (section 6).

2. Existing Work

A number of methods have been developed to analyse, capture and cluster businesses. This section will consider methods that have been developed to classify companies, with some degree of automation, using publicly available data. The methods presented in this section rely on varying degrees of manual analysis from the user.

Growth Intelligence, a UK company, have constructed a dataset that contains data relating to a cross-section of companies, registered at Company House, that were active in the UK up until August 2012. They use their own software to 'enrich' the data according to the digital signatures of the companies. What makes up a digital signature will depend on the company but could include different public data sources (e.g., company's webpage, patent applications, social media, search engine traffic). The dataset is combined with text analytics and fed into their machine learning tools to produce a classification of the companies. This could be according to sector or products. The method allows the company to be regularly reclassified thus allowing for the dynamic nature of the field. Machine learning allows correlations to be identified and classifications made in real-time.

The National Institute for Economic and Social Research (NIESR) has published a report that measures the size of the Digital Economy by adopting a method that makes use of the data from Growth Intelligence [2]. Companies are considered to exist in the Digital Economy if they exist in a digital sector or if their outputs could be considered digital. To identify these companies, the method presented in [2], is applied to a sample dataset containing 1.868 million UK companies. The method involves six steps: (1) a shortlist of sectors and product groups (21 and 15 respectively) are produced by taking the sub-set of companies who have 'Digital Economy SIC codes' and the relevant sectors and product classifications given by Growth Intelligence; (2) companies that have minimal presence in the Digital Economy are excluded with the use of a predetermined threshold; (3-4) the sector and product/services lists provided by Growth Intelligence are manually edited to remove any sector groups or products/services that are considered irrelevant (based on predefined rules); (5) a precision check is completed by analysing the Growth Intelligence product and sector groups; and, (6) the sample is cleaned according to 'sector-by-product groups'.

After applying their method, NIESR give a conservative estimate of 269,695 companies (14.4% of all companies in 2012), and a more generous estimate of 471,120 companies making up the Digital Economy. Both these figures are greatly lower than the 167,000 (10.0% of all companies) given by the government, at the time of the report who relied on traditional SIC codes.

Neely (2009) presents a method to quantify the extent of servitization in manufacturing. To apply the method, manufacturing companies, identified via their Standard Industrial Classification SIC codes, who had over 100 employees were extracted from the OSIRIS database [15]. Companies identified as having either the wrong SIC code (i.e., not a manufacturing company), no description, declared bankruptcy or classified as pure service were omitted from the analysis leaving a total of 10,634 companies from 25 countries.

The data was manually coded to allow companies to be classified as 'pure manufacturers', 'servitude

manufacturers' or 'pure service'. The coding was based on the description that was given by the company. Keywords from the first 50 companies were manually identified that would allow a classification to be made and a codebook produced. 12 services were identified as being frequently mentioned within the codebook. The terms and phrases associated with the 12 services, along with their variants, were used to code the 50 companies with the use of Excel formulae.

The original manual coding and the 'automatic' coding via Excel were compared and moderations made to the Excel search functions to allow for any errors and discrepancies that had been found to be processed. All companies within the dataset were classified with the use of a conservative approach: unless there was clear evidence, a firm was classified as pure manufacturing. A random sample of the classified firms were reviewed to ensure that there had been no 'significant miscodings'. The analysis by Neely showed that 30.05% of the companies had servitized.

Wu, et al. [16] recognises that data held on webpages provides a rich dataset that can be explored to extract information about the Digital Economy and presents a semi-automatic method that makes use of this data to produce a taxonomy of businesses. The basic level concepts of a business model are described using a core ontology: a tuple F that contains the users (U), tags (T), resources (R) and the tertiary relation (Y) that exists over U, T and R. Given new data, new concepts and relations can be identified. Data is scraped from webpages of companies who have a digital focus. The data is processed using text analytics (a Stanford Parser) to identify the semantic units of each sentence. Supervised learning can then be used on the parsed text to produce the taxonomy based on the core ontology. The aim is to learn the taxonomy of concepts O which contains the concepts C, properties P, instances I and the set of rules S that relate C, P and I.

3. The proposed method

The proposed method comprises three distinct components: web scraping; machine learning; and taxonomy generation. Each operates as a "black box", requiring only the output from the previous step. This loosely coupled architecture provides for flexibility in the tool, as the tool develops more advanced replacements can easily be substituted. This section will describe each component in turn.

3.1. Web scraping

The web scraping component in the proposed method operates by identifying a pre-defined list of parameters in a given set of web pages. The procedure for scraping a web page is specified in Algorithm 1. Each website is examined for the presence of each parameter. A binary string, where each bit (i.e., character) represents one of the parameters, is used to represent each website; if a parameter is found, then the appropriate parameter bit is set to 1, otherwise it is set to 0. Both lists (parameters and websites) are held externally and used as inputs to the web scraping component, so the method can be easily extended to additional parameters and web sites; this maximises the flexibility of the overall method.

Algorithm 1 Web Scraping

```
Read the list of web sites (W)
Read the list of parameters (P)
for each website w do
    for each parameter p do
        if website w contains parameter p then
            Let b_wp = 1
        else
            Let b_wp = 0
        end if
    end for
end for
```

The pre-defined list of parameters is chosen to capture different business models. In its simplest form a business model can be characterised by three concepts: the value proposition, value creation and worth capture [5]. Value proposition is the system of resources employed to deliver the purpose of the enterprise, which may be a product, service or a combination of both. Value creation is the realisation of the value proposition for the benefit of a customer and occurs in a specific context; as context changes the determination of value realised is subject to change [17]. Worth capture is the ability of all parties to capture benefit and is usually measured in terms of monetary exchange. This work will focus on worth capture in terms of monetary exchange but further work may also include other worth capture such as goodwill etc. [18].

For each of the three areas a set of relevant questions were considered to help identify an initial set of suitable keywords (i.e., the keywords were those that would allow these questions to be answered). The validity of these keywords would be explored through application to a sample dataset. Value proposition (vp) considered the questions *'what do we offer?'* and *'what is the customer value?'*. Value creation (vc) focused on *'what are the key activities?'*. Worth capture (wc) considers *'what is the market value, revenue or cost structure?'*. The initial set of keywords resulting from

these questions can be found in table 1; these become the parameters for the web scraping. For initial proof of concept the number of keywords was kept to a minimum, but the method allows for keywords be expanded upon in later work and synonyms added. In future versions of the method, the parameters should be automatically identified, instead of being manually defined (see section 6).

Table 1. Parameters to identify different business models

Concept	Keywords
vp	product/service/review
vc	aim/purpose/function/ objective/goal/mission
wc	cost/price/subscription/ monthly

3.2. Clustering

Clustering is a process where similar data points are identified as those with similar characteristics. In this case, the websites are the data points and the presence or absence of parameters on those pages are the characteristics. The aim of the clustering component is to find clusters that correspond to different business models that can be defined according to the combination of the parameters.

Most clustering methods require the definition of a distance between data points. The method presented here employs the Hamming distance (Hamming, 1950), which computes the distance between data points in terms of the number of common pairwise elements. Here, this computes a distance in terms of the frequency that two websites both feature each parameter. Two websites with most parameters in common are "close", while two with few shared parameters are "distant".

There are many different types of clustering methods available (e.g., partitioned, hierarchical, model-based) [19]. The proposed method is being developed for unlabelled data and, therefore, an unsupervised approach is needed. The k-means clustering algorithm [20] is used to cluster the websites based on the parameters that have been identified. K-means is a well-known clustering algorithm used to partition data sets by placing k points into the data, each of which represents a cluster. Following an iterative procedure, the data points are assigned to the cluster to which they are nearest. The k cluster points are then adjusted so that they correspond to the average values of the data points in their cluster. Over time, this moves the cluster points so that they accurately represent the clusters present in the data. K-means was chosen as the initial clustering algorithm to explore the potential of the proposed method since it can be considered to have low-time complexity but typically high-computing efficiency [19]. Further work would include an investigation and comparison of alternative clustering algorithms, especially for when more input parameters and websites need to be considered (see section 6).

The method of k-means clustering can suffer from the "curse of dimensionality" [21], whereby it is difficult to identify clusters within high-dimensional data. For the method presented in this paper each parameter will represent a dimension and thus be dealing with high-dimensional data. To address this problem, the dimensionality of the data (i.e., the parameters) was reduced using MultiDimensional Scaling (MDS) [22], which constructs new low-dimensional data points to represent the original data. Two data points in the low dimensional space are close if the corresponding high dimensional points were close. Conversely, if the original points were far apart, then their low dimensional equivalents are also distant. The clustering algorithm can then operate in the low dimensional space, and is likely to be more successful. A positive side effect of this approach is the clusters can be easily visualised. Once the dimensionality has been reduced with MDS, the Euclidean distance between the low dimensional points is used in combination with k-means clustering to identify the clusters present within the data.

An additional problem with the k-means approach is that the value of k must be specified prior to clustering. A difficulty with this is that this value is rarely known for real datasets. Fortunately, the actual clustering process is computationally inexpensive and can be repeated many times. As such, the process is run for values of k between 2 and 20 and the 'quality' of the clustering evaluated to see what the best setting of k is. The silhouette coefficient [23] is used to determine cluster quality; this measure considers the distance between a data point and the other data points within the same class, as well as the distances between the data point and all data points within the next nearest cluster. Since k-means is a stochastic algorithm and the same result is unlikely between runs, the process was repeated to determine the likelihood that the clustering quality was correct.

Figure 1, shows that error bars become very tight as the number of clusters increases, indicating that across all 30 runs the result was within a small region and is reliable. An upper value of 20 is justified given that this is where the silhouette coefficient levels off. The tightness of the error bars within figure 1 also justifies the use of 30 runs. In this initial work the value of k has been chosen manually following the silhouette

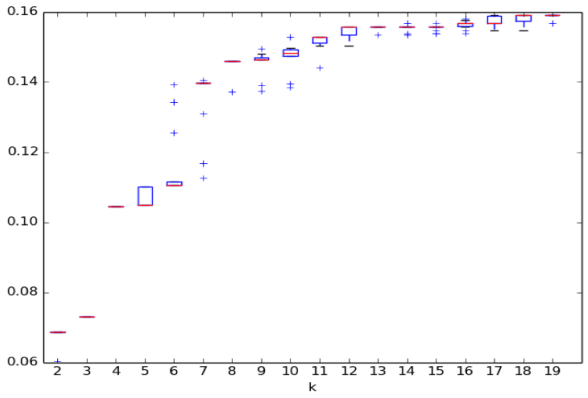

Figure 1. Clustering quality for various values of k. The value of k chosen trades off the number of clusters (to be minimised) against the quality of the clustering, which ranges from -1 to 1 (to be maximised). In this case, the value k=8 is chosen

experiments. This could, and should, be automated and is an area that has been highlighted for future work (see section 6).

3.3. Taxonomy generation

Previous methods to classify the nature of businesses have resulted in frameworks and canvases being produced that highlight the key concepts that can describe and differentiate business models (e.g., [3, 24]). Currently the simplest business model framework [5] was used to identify the parameters that might highlight the business model that a company uses. In later work it will be possible to use more complicated models.

In ontological terms, a taxonomy is 'a hierarchy consisting of terms denoting types (or universals or classes) linked by subtype relations' [25]. The common features of what is being analysed allows things to be grouped into types where a type is an entity in the domain that you are considering. For the method presented here, we wish to produce a taxonomy of businesses based on their business model where the business model is determined by the combination of parameters that have been registered. The method constructs the taxonomy according to the outputs of the clustering process outlined in section 3.2 and the combination of parameters that have allowed them to be clustered. Therefore, for each node in the taxonomy, a set of companies will exist for which the combination of keywords (represented by that node and its parents) have been identified.

In addition to the generation of a taxonomy of business models the method could be used to measure activity from product to service over time (i.e., servitization). To measure servitization the method would categorise companies according to the degree of product/service offer. As an initial step to this, the categorisation will be on the presence of keywords particularly product and service; once the method has been refined, combinations of parameters could be a better measure the degree of servitization and this is left for further work (section 6).

4. Application

This section explores the potential of the proposed method by applying it to a sample subset of websites: the list of the West Country's 150 biggest businesses as listed by Western Daily Press [14]. This set of webpages were chosen since it contained a variety of businesses in different sectors. This case study can be considered a proof of concept , which will highlight the method's strengths and areas that need improving.

The method has been implemented in Python and relies on several open source Python modules; these are discussed as their use arises. Out of these 150 businesses, 109 had a website that the web scraping component could access. The other 41 websites either had security that would not allow access to the website content or did not have a valid URL. Two sets of results are presented to highlight the use of different input parameters to the web scraping component: the 13 parameters listed in table 1, and an extended parameter list.

4.1. Input of 13 parameters

When applying the 13 parameters given in table 1 to the 109 webpages, 39 webpages returned a '0' for all parameters (i.e., none of the parameters were found). Table 2 shows the number of websites where each parameter had been found (i.e., a 1 was returned). 18 of the websites registered both 'product' and 'service' as a parameter. Only one parameter was registered for 32 of the webpages but the parameter varied between the various websites. 20 and 14 webpages registered 2 and 3 parameters respectively. Four and five parameters were found on 2 webpages. No website registered all the parameters.

In total 8 clusters were identified (denoted C0 to C7). Figure 2 shows the results of applying the clustering method, outlined in section 3.2, on the output of the web scraping when using the parameters found in table 1 as the input. Each of the 109 websites were classified according to the combination of parameters that had been registered.

Initially whether the parameters product or service had been found were used as a simple mechanism to

Table 2. Number of websites registering each parameter

Parameter	No. of Websites
Function	0
Monthly	1
Review	1
Goal	2
Mission	3
Purpose	3
Objective	4
Aim	8
Cost	8
Subscription	11
Price	13
Service	37
Product	39

identify servitization and the degree of product/service offer. These two parameters seem very important for clustering when building the taxonomy. It appears that businesses are initially classified according to whether the parameters product and service are present within the website. The only other parameter that occurs throughout the companies in a cluster is price, if both product and service have been identified (i.e., C7). The remaining clusters seem to focus on how many of the other parameters have been found; there does not seem to be any pattern in the combinations of the remaining parameters once product, service and price have been taken out.

A taxonomy based on the output of the clustering could initially classify between whether the parameters product and service have been found whether singularly (i.e., just product or service) or both had been identified. Further classification could be whether the parameter price had been found followed by how many parameters have been identified. It is difficult to classify according the remaining parameters since no patterns were evident from the clustering. This leaves a single tiered taxonomy that is not very useful and does not capture fine level detail.

C0 Neither product or service parameters identified. A total of 12 companies. All 12 companies registered at least one parameter (review, purpose, objective, cost, price or subscription). One company registered two parameters: subscription and review. The companies from this cluster came from different sectors including car dealerships, energy, travel, media, mining, technology and retail.

C1 Product and service parameters identified. All 15 companies registered the parameters product and service. Some companies registered one parameter (aim, purpose, objective, cost, price, or subscription) in addition to product and service. The companies from this cluster came from different sectors including security, engineering, oil and gas, telecommunications, manufacturing, insurance and health.

C2 Product parameter only. All 13 companies only registered the parameter product. The companies from this cluster came from different sectors including financial services, food, insurance, retail, technology and manufacturing.

C3 Neither product or service parameters identified. 39 companies where none of the parameters had been located on their website. The companies from this cluster come from different sectors including retail, manufacturing, engineering, financial services, entertainment, truck dealer and recruitment.

C4 Service parameter only. All 10 companies registered service, and not product, with at least one other parameter registered (aim, goal, cost, price or subscription). The maximum number of parameters registered by any company in this cluster was 4. The companies from this cluster comes from different sectors including energy, legal, construction, financial services, logistics and car dealerships.

C5 Service parameter only. All 10 companies registered service and not product with no other parameters registered. The companies from this cluster comes from different sectors including construction, accident management, financial services, food and water.

C6 Product and service parameters identified. All 8 companies registered product and service with at least one other parameter registered (aim, purpose, mission, cost, price, subscription or monthly). The maximum number of parameters registered by any company in this cluster was 3. The companies from this cluster comes from different sectors including suppliers, manufacturing, technology and insurance.

C7 Product and service parameters identified. Both companies registered product, service and price. Each company also registered at least one additional parameter (objective, goal, mission, cost, or subscription). The companies from

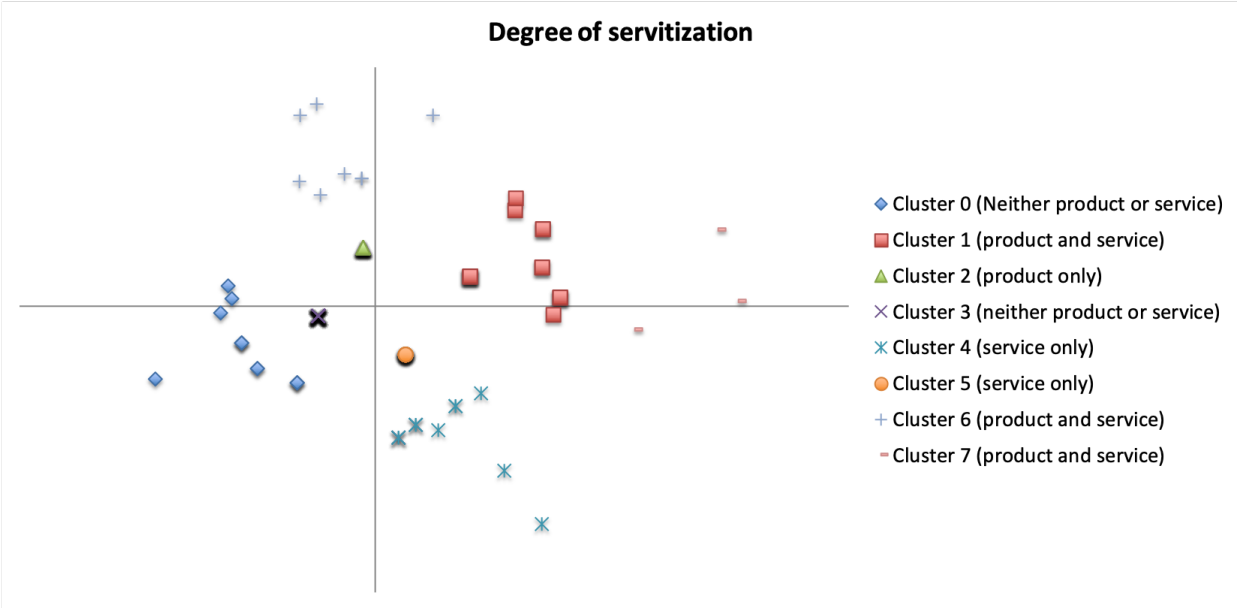

Figure 2. Company websites clustered according to 13 parameters found during web scraping. Some companies have been overlaid due to extremely close similarity

this cluster comes from three different sectors: technology, manufacturing and oil/gas.

4.2. Extending the input list

The list of 13 parameters resulted in 39 companies not registering any of the parameters. The lack of taxonomic structure after considering the parameters product and service suggests the need for parameters that would allow for more subtle distinctions to be made. The sentences where the original 13 parameters had been extracted were analysed to see if they could allow better insight for clustering. Each sentence was decomposed into its components (i.e., the words that is comprises) and each of these words were then used to form an extended parameter list. It was hoped that this method might identify some context without extended semantic analysis.

After completing the process for all 109 websites, the extended parameter list comprised 1,754 parameters. Some words were thought to be unsuitable for classifying business models (e.g., connectives, articles, pronouns, numbers and non-English words). These were removed leaving a parameter list of 1,191 words to feed into clustering. To give an indication of the additional words that were found, figure 3 depicts the list of parameters as a word cloud where the words that were registered the most appearing more prominently.

Using the extended word list as an input to the web scraping resulted in the 8 clusters. When constructing the taxonomy according to the combination of parameters that have allowed them to be clustered, there are minimal patterns evident from the clustering. The initial distinction when considering only a reduced parameter list looked at the parameters product and service. However, this distinction cannot be made with the extended list since not all members in many of the clusters all register one, both or neither. These results show that simply extending the original parameter list in this way does not help measure the degree of servitization.

5. Discussion

The results shown in section 4 show that the proposed method has potential and illustrate how the proposed method can classify companies from real-world data obtained from their own websites. Although some insight can be gathered with regards to the clusters that have been generated, it is clear that the way in which the parameters are identified needs advancing.

The clusters that have been produced with the initial 13 parameter list (figure 2) show only one cluster (C2) is considered product only, two clusters are service only (C4, C5) and three clusters a combination of product and service. Parameters need to be identified that would help clarify more detailed distinctions in terms of degrees of servitization. However, the results also show that just considering the words around the 13 parameters is not

Figure 3. Word cloud showing frequency of words registered by websites in the extended word list.

sufficient.

Both variations of running the method (i.e., with the original parameter list and with the extended parameter list) have each produced a set of 8 clusters. However, how the companies are clustered in each set varies greatly. Each set of clusters has one cluster that is much larger than the rest: C0r (C0 for the reduced parameter list) and C3e (C3 for the extended parameter list). Only 19 companies are shared between these two clusters. The remaining companies from C0r have been split between the other six clusters (C1e, C2e, C3e, C4e, C5e, C6e, C7e). There seems to no significant group of companies that have remained in the same cluster when comparing the two variations of the method. This suggests that the additional parameters that have been included within the extended parameter list has greatly affected the results but not helped measure the degree of servitization.

The inclusion of additional parameters has reduced the number of companies that could not be classified. However, the extended parameter list has made it more difficult to identify the taxonomic patterns than when there were only 13 parameters. The generation of the taxonomy relies on the combinations of the parameters that have been found in each cluster. The lack of evident patterns in the clustering has made the generation of the taxonomy difficult. Thus, emphasising the importance of selecting the correct parameters and that the method still does not have the optimum set of parameters for capturing business models and servitization. Many of the parameters have only been identified by one of the companies in each cluster. Further work is needed to identify the core combinations of parameters that can distinguish the different business models.

A simple approach is currently adopted to measure servitization by considering the presence of the keywords product and service. It is possible that combinations of keywords could better identify servitization; different combinations of keywords might also suggest different degrees of servitization. Further work is needed to establish these combinations. It is possible that the identification of the parameters that

answer the relevant questions to value capture, value creation and worth capture might require more advanced text analytics of the sentences where keywords such as aim and objective are found.

With both variations of the method presented in section 4, each cluster represents companies from different sectors. Since the clustering is done on the presence of keywords this is not surprising. The clustering could be considered to cluster companies according to how they describe themselves. The combinations of different sectors within each cluster could provide interesting insight into marketing strategies and overlaps in objectives of organisations within different sectors. Such information could help companies identify any encroachment or how existing companies are positioning themselves in terms of market strategies (e.g., companies could compare their use of subscriptions, under worth capture, to that of their competitors to understand the use of a different payment model).

The analysis is not limited to firms who meet the necessary qualification to appear in a database, such as geography, turnover, registration etc. This could enable companies to identify new or potential competitors. The ability to repeat the method over time to understand business model trends could help companies compare themselves to their competitors.

6. Further Work

A set of extensions to the method have been identified as areas for future work particularly around parameter identification, clustering, the automatic generation of the taxonomy and identifying additional datasets to which the method could be applied.

An analysis of the results presented in section 4 has shown that the choice of parameters is very important. The success of the whole process is largely driven by the presence of certain parameters in company websites. The use of synonyms is therefore problematic, as a company using a synonym to a parameter within their website will cause that parameter to be registered as not present. More advanced text analysis of the website content, for example including synonyms in parameter lists and considering words around found parameters, is recommended.

Further work is needed to establish which combinations of parameters could be used to classify business models and measure the degree of servitization. Currently the list of parameters must be pre-defined. Further work should be undertaken to see if parameters could be automatically identified from company websites (e.g., considering menu items listed). Methods to validate the usefulness of keywords must also be developed.

More advanced feature engineering based on web scraping and inclusion of data from social media could be included in the method thus allowing the different business models to be identified more accurately. The security restrictions on certain websites prevented the scraping tool from accessing them and 19 companies out of the 150 chosen did not have a listed website. Alternative sources of data, such as social media channels representing those companies, may provide a viable approach to including them. These additional datasets may also allow us to focus on different aspects of the business model (e.g., value proposition and customer value might be established from online reviews and/or social media).

The method as presented in this paper is not fully automated since the number of clusters must be predefined. Work has begun on refining the machine learning component to allow the method to be fully automated. While k-means has shown to provide some interesting clusters in the case study provided, as more companies and parameters are included the suitability of the algorithm may be reduced. Alternative clustering algorithms need to be considered (vector machines and self-organising maps are likely candidates). A comparison of those clustering algorithms for the proposed method would also need to be considered. Once the web-scraping and machine learning components have been refined, the taxonomy generation could also be enhanced to include more ontological theory and the automatic generation of the taxonomy.

The method is very flexible and could be used to classify companies according to many different features with the choice of parameters reflecting the feature that is to be studied (e.g., servitization). If the method is to be used to identify companies within the Digital Economy a set of keywords is needed that would reflect the Digital Economy. Work as begun on identifying this list. The work is focused on developing a method that allows businesses to be classified based on the business models that they operate. Focus has not been placed on what may constitute a business that is part of the Digital Economy due to the lack of agreed definition of what should be included.

Once refined the method should be evaluated against existing methods (such as those listed in section 2) and the SIC codes of the companies that are being clustered. This will help to establish the validity of the proposed method.

7. Conclusion

A flexible method, using open source tools, has been developed that allows businesses to be classified according to a set of parameters. The method comprises three distinct components: a tool for scraping content from websites; a clustering mechanism; and an approach for generating a taxonomy from the clusters identified. Each operates as a "black box", requiring no interaction with the previous step but relying simply on its output. Such a taxonomy may lead to classifications that characterise the extent of Digital infusion into business models. In developing the method we have identified a number of areas that need further consideration particularly the choice of parameters and the requirement for additional data sources other than a company's website.

8. Acknowledgement

This work was supported by a grant from NEMODE – New Economic Models in the Digital Economy, Network+ EPSRC funded programme, as part of the project "Using computational methods to produce a taxonomy of business models of the digital economy" (2016).

References

[1] T. Baines, H. Lightfoot, S. Evans, A. Neely, R. Greenough, and e. a. Peppard, J., "State-of-the-art in product-service systems.," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 221, no. 10, pp. 1543–1552, 2007.

[2] N. Max, A. Rosso, T. Gatten, P. Majmudar, and A. Mitchell, "Measuring the uk's digital economy with big data," tech. rep., National Institute of Economic and Social Research (NIESR), 2013.

[3] A. Osterwalder, Y. Pigneur, and T. Clark, *Business model generation: A handbook for visionaries, game changers, and challengers.* Hoboken, NJ: Wiley, 2010.

[4] C. Baden-Fuller and V. Mangematin, "Business models: A challenging agenda," *Strategic Organization*, vol. 11, no. 418 - 427, 2013.

[5] G. Parry and P. Tasker, "Value and servitization; creating complex deployed responsive services?," *Strategic Change*, vol. 23, pp. 303–315, 2014.

[6] S. Vandermerwe and J. Rada, "Servitization of business: adding value by adding services," *European management journal*, vol. 6, no. 4, pp. 314–324, 1988.

[7] T. Baines, H. Lightfoot, O. Benedettini, and K. J.M, "The servitization of manufacturing. a review of literature and reflection on future challenges,," *Journal of Manufacturing Technology Management*, vol. 20, no. 5, pp. 547–567, 2009.

[8] R. Oliva and R. Kallenberg, "Managing the transition from products to services," *International Journal of Service Industry Management*, vol. 14, no. 2, pp. 160–172, 2003.

[9] A. Tukker, "Eight types of product-service system: Eight ways to sustainability? experiences from suspronet.," *Business Strategy and the Environment*, vol. 13, no. 4, pp. 246–260, 2004.

[10] E. Fang, R. Palmatier, and J. Steenkamp, "Effect of service transition – strategies on firm value," *Journal of Marketing*, vol. 72, pp. 1–14, 2008.

[11] A. Neely, "Exploring the financial consequences of the servitization of manufacturing," *Operations Management Research*, vol. 1, pp. 103–118, 2008.

[12] A. Eggert, J. Hogreve, W. Ulaga, and E. Muenkhoff, "Industrial services, product innovations, and firm profitability: a multiple-group latent curve analysis.," *Industrial Marketing Management*, vol. 40, no. 5, pp. 661–670, 2011.

[13] I. Visnjic-Kastalli and B. Van Looy, "Servitization: Disentangling the impact of service business model innovation on manufacturing firm performance," *Journal of Operations Management*, vol. 31, no. 4, pp. 169–180, 2013.

[14] W. D. Press, "Top 150 businesses guide 2014. south west business.."

[15] B. V. DIJK, *OSIRIS*. International: Beureau van Dijk Electronic Publishing., 2012.

[16] C. Wu, Y. Cai, M. Zhao, S. Huang, and Y. Guo, "Generating computational taxonomy for business models of the digital economy," in *Database Systems for Advanced Applications (DASFAA 2016)* (G. H., K. J., and S. Y., eds.), vol. 9645 of *Lecture Notes in Computer Science*, pp. 126 – 133, Springer, Cham, 2016.

[17] C. Zott, "Dynamic capabilities and the emergence of intra-industry differential firm performance: Insights from a simulation study," *Strategic Management Journal*, vol. 24, pp. 97–125, 2003.

[18] D. Lepak, K. G. Smith, and M. S. Taylor, "Value creation and value capture: A multilevel perspective," *Academy of Management Review*, vol. 32, no. 1, pp. 180–194, 2007.

[19] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms.," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[20] J. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[21] R. E. Bellman, *Dynamic programming*. Princeton, NJ.: Princeton University Press, 1957.

[22] T. F. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall., 2001.

[23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 2987.

[24] C. B. Mangematin, "Business models: A challenging agenda," *Strategic Organization*, vol. 11, pp. 418–427., 2013.

[25] R. Arp, B. Smith, and D. S. Andrew, *Building ontologies with Basic Formal Ontology.* Massachusetts: MIT Press, 2015.