Olatunji, I.E., See-To, W.K.E. & Papagiannidis, S. (2019). Longitudinal analysis of economic clusters: A novel methodology and application of UK regions. In *Proceedings of the 19th International Conference on Electronic Business* (pp. 570-572). ICEB, Newcastle upon Tyne, U.K., December 8-12.

Longitudinal Analysis of Economic Clusters: A Novel Methodology and Application of UK Regions

(Abstract Only)

Iyiola E. Olatunji*, The Chinese University of Hong Kong, iyiola@link.cuhk.edu.hk
Eric W.K. See-To, Faculty of Business, Lingnan University, Hong Kong, ericseeto@ln.edu.hk
Savvas Papagiannidis, Business School, Newcastle University, Newcastle upon Tyne, UK
savvas.papagiannidis@newcastle.ac.uk

ABSTRACT

Standard Industrial Classification (SIC) classify organizations based on their business activities. However, choosing appropriate SIC code that represents an organization's business activities in a challenging task. In the UK, there are almost 100 categories each having several subcategories of predefined business activities designed by experts. However, such scheme cannot cater for emerging business needs while some organizations cannot be easily defined by a single SIC code, due to the complexity of their business nature. Similarly, if a company expands or changes its operation during the year, a new SIC code needs to be assigned. This results in organizations having difficulties picking representative SIC code to use in defining their business activities. In this paper, we propose a dynamic framework that can automatically group organizations based on their business activities. Our framework leverages techniques from topic modelling. Result shows that our proposed framework can automatically adapt to changing business needs and cluster organizations effectively.

Keywords: Big data analytics, longitudinal analysis, standard industrial classification, topic modelling.

*Corresponding author

INTRODUCTION

Accurately classifying industries into clusters based on their business activities is essential in order to provide valid statistical information and to quantify the impact of industry structures on innovations and investments. These classification systems provide insight into the economy and trade and it enlightens policy makers on current industrial activities. In the UK, the standard industrial classification (SIC) is used for classifying economic activities (Smith & James, 2017).

SIC consist of 99 categories and several subcategories used to define the business activities of organizations in the UK. The design of the SIC system reflected the way industry structures were understood by policy makers at the time creation (Porter, 1981). The duty of the policy makers is to promote equilibrium in the economy. Therefore, the SIC system is designed based on the past business activities of organizations. For example, within the "Agriculture, Forestry and Fishing" group, there are 40 different SIC codes ranging from growing of cereals to freshwater aquaculture.

However, such classification schemes do not adapt to the evolving changes in business activities. For example, changes in the standard industrial classification have occurred approximately every 10-15 years in the UK. Similarly, when such changes occur, they require large amounts of resources to implement and can create disruption in time series (Brook, Matthews, & Darke, 2011).

LITERATURE REVIEW

In the UK, SIC has been revised seven times since 1948. These revisions, with the latest in 2007, have failed to capture recent business activities (Hughes & Brook, 2009). Business activities not explicitly captured by the classification codes are grouped into the "other" categories (e.g other mining and quarrying category) (Hrazdil & Zhang, 2012). Similarly, the revision of SIC has brought about differences in codes across databases, thereby causing inconsistences and failure to identify industries that have similar operating characteristics (Hrazdil, Trottier, & Zhang, 2014).

Also, products and services have become more complex as a result of changes in innovation and technology (Dalziel, 2007). These technological changes may affect production processes and appropriate classifications because products and services depend on a wide range of supporting products and services to function fully.

Lastly, the development of the SIC system indicates the perception of the industry structure as understood by the experts as at the time of its creation (Hrazdil, Trottier, & Zhang, 2013). This creates an inflexible and rigid classification scheme as opposed to a dynamic scheme that can instantly capture emerging business activities.

METHODOLOGY

To solve the above problems, we propose a dynamic framework based on techniques from topic modelling that is robust to emerging business needs and can automatically assign a cluster to an industry based on their business activities. Topic modelling is a method of discovering group of words also known as topic from a corpus. It is a type of statistical model can be used to identify recurring patterns of words from text and discover hidden semantic structures embedded in text. Topic modeling techniques have proven to be effective in various natural language modelling tasks such as semantic mining and discovery of latent topics in documents. In topic modelling, a topic is defined as a list of co-occurring words with statistical significance. A document in our case is the textual description of firm's activities and corpus is the collection of documents. Techniques used in topic modelling includes Explicit Semantic Analysis (ESA) (Gabrilovich, Markovitch, & others, 2007), Latent semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 2001), Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) and Hierarchical Dirichlet process (HDP) (Teh, Jordan, Beal, & Blei, 2005). Of all the above techniques used for topic modelling, Latent Dirichlet Allocation (LDA) is most suitable for our task and performs better than others.

Therefore, our framework utilizes Latent Dirichlet Allocation (LDA), a generative statistical model that can automatically extract topics from documents (Blei, Ng, & Jordan, 2003). In LDA, each cluster, also known as a topic, consists of a series of words, each with a weight that indicates the importance of the word in the cluster. The aim of LDA is to infer latent topics from the corpus. LDA assigns a set of topics for each document with dirichlet distribution and computes a probability over the words in the entire corpus. Then for each document, LDA calculates the probability of the words in the documents belonging to the mixture of a word set of the topics.

Let C be a corpus consisting of D documents (description of firm's activities) and T be the number of topics. Each document $d \in D$ consists of k words $\mathbf{w} = (w_1, w_2, \dots, w_k)$. For each document $d \in \{1, \dots, D\}$, we choose a multinomial distribution θ sampled from a Dirichlet distribution with $\propto$. Similarly, for each topic $t_n \in \{1, \dots, T\}$, we choose a multinomial distribution φ sampled from a Dirichlet distribution with β. Then each word w_k in document d is sampled conditioned on the t_nth topic from the multinomial distribution φ.

In the generative process described above, $\propto$ and β are hyper parameters, ϑ and φ are latent variables and words w_k in the document are the observed variables. We compute the probability of the document over the corpus and infer latent variable and hyper parameters by:

$$p(C|\propto,\beta) = \prod_{d=1}^{D} \int p(\theta|\alpha) \left(\prod_{n=1}^{k} \sum_{t_n=1}^{T} p(t_n|\theta) p(w_k|t_n,\beta)\right) d\theta \quad (1)$$

Where $p(t_n|\theta)$ is a multinomial characterized by θ, $p(w_k|t_n,\beta)$ is the a multinomial over the words and $p(\theta|\alpha)$ is the Dirichlet. As shown in the equation above, the Dirichlet is sampled for each document and topics are sampled repeatedly within the document. Since LDA treats documents as a mixture of topics (where a topic is the probability distribution over the set of words), the Dirichlet $p(\theta|\alpha)$ gives the mixture weight.

Our dataset consists of information extracted from the websites of over 14,000 companies from 2000 to 2019 in the UK across six regions (Guernsey, Ireland, Northern Ireland, Wales, East London and Scotland). This information includes the description of the firm's activities, number of employees, turnover, SIC codes and address. Precisely, the description of the firm's activities was treated as our text corpus for performing the analysis. The aim of our analysis is to group companies into clusters based on their business activities. To estimate the parameters of our LDA model and to infer the distribution of topic and words, we used Variational Bayes (VB) inference. VB uses parametric approximation to the posterior distribution of parameters α and β, and latent variables θ and φ. VB allows for faster convergence and iteration can be computed faster than Gibbs sampling method. Based on experiments, we set $\beta = 0.01$ while we let α automatically learn asymmetric prior from the corpus.

We split the textual description of the firm's activities into sentences and sentences into words. We performed stemming, lemmatization, and removed punctuation and stop words. To extract reasonable topics, our model considers the frequency of words within a topic and also across topics to prevent certain words appearing in all the topics. We performed both cross-sectional analysis and longitudinal analysis. For our cross-sectional analysis, we cluster companies based on their business activities within their regions. We determined the optimal number of clusters based on the data about that region. For the longitudinal analysis, we consider a triplet consisting of region, quarter and year. For each triplet, we performed coherence analysis to determine the optimal number of clusters for the business activities. We use the above triplet as input into our LDA model to show the evolution of clusters for each <region, quarter, year> triplet.

RESULT

Our preliminary result shows that our proposed framework can effectively cluster new business activities and adapt the number of clusters dynamically based on the textual description of the business need. Secondly, the implementation of our framework provides a low-cost solution that can detect homogeneity in business activities. Lastly, our result shows that our model reduces within-class heterogeneity and is robust to changes in technology and business needs.

ACKNOWLEDGMENT

This work is supported by the Office of National Statistics, UK.

REFERENCE

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Brook, K., Matthews, D., & Darke, J. (2011). Changes to the picture of the UK economy - impact of the new SIC 2007 industry classification. *Economic & Labour Market Review*, *5*(3), 41–61. http://doi.org/10.1057/elmr.2011.30

Dalziel, M. (2007). A systems-based approach to industry classification. *Research Policy*, *36*(10), 1559–1574. http://doi.org/https://doi.org/10.1016/j.respol.2007.06.008

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407

Gabrilovich, E., Markovitch, S., & others. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI* (Vol. 7, pp. 1606–1611)

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1–2), 177–196.

Hrazdil, K., Trottier, K., & Zhang, R. (2013). A comparison of industry classification schemes: A large sample study. *Economics Letters*, *118*(1), 77–80. http://doi.org/https://doi.org/10.1016/j.econlet.2012.09.022

Hrazdil, K., Trottier, K., & Zhang, R. (2014). An intra- and inter-industry evaluation of three classification schemes common in capital market research. *Applied Economics*, *46*(17), 2021–2033. http://doi.org/10.1080/00036846.2014.892200

Hrazdil, K., & Zhang, R. (2012). The importance of industry classification in estimating concentration ratios. *Economics Letters*, *114*(2), 224–227. http://doi.org/https://doi.org/10.1016/j.econlet.2011.10.001

Hughes, J., & Brook, K. (2009). Implementation of SIC 2007 across the Government Statistical Service (GSS). *Economic & Labour Market Review*, *3*(8), 67–69. http://doi.org/10.1057/elmr.2009.145

Porter, M. (1981). The Contributions of Industrial Organization to Strategic Management. *The Academy of Management Review, 6*(4), 609-620. Retrieved from www.jstor.org/stable/257639

Smith, P. A., & James, G. G. (2017). Changing Industrial Classification to SIC (2007) at the UK Office for National Statistics. *Journal of Official Statistics*, *33*(1)

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (pp. 1385–1392).