SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Adopting Text Mining for Patent Analysis to Determine the Attribute and Segment in Automotive Industries

Amir Syafiq Syamin Syah Amir Hamzah[1,3,*], Hafizah Farhah Saipan@Saipol[1,3], Syarifah Zyurina Nordin[1,3], Zatul Alwani Shaffiei[2], Naoki Ohshima[4]

[1] Department of Management of Technology (MoT), Malaysia–Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[2] Department of Electronic Systems Engineering (ESE), Malaysia–Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[3] Intellectual Property and Innovation Management (IPIM) iKohza, Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[4] Graduate School of Innovation and Technology Management, Yamaguchi University, Yamaguchi, Japan

ARTICLE INFO

Article history:
Received 17 June 2023
Received in revised form 24 October 2023
Accepted 10 November 2023
Available online 16 January 2024

Keywords:
Text mining; word cloud; co-occurrence networks; correspondence analysis; NLP; pre-processing

ABSTRACT

Analysing massive patent documents in heavy industries including automotive has become important in recent years as they contain a lot of information that is extremely difficult to deal with from huge numbers and various forms. Important documents such as patent data on a huge scale has become a major concern owing to time constraints and enormous costly work. In natural language processing (NLP), text mining is used to determine several features such as segmentations and attributes of data. The important step in data mining is pre-processing data information from massive text data. In this paper, the fundamental concept of pre-processing and data analysing is examined to provide accurate and meaningful information from the chosen data set. This study focuses on two automotive companies, namely Mazda and Mitsubishi to describe the similarities, distances and frequencies between several patent documents. To demonstrate the behaviour of selected patent documents, word cloud, co-occurrence networks and correspondence analysis are also presented.

1. Introduction

Automotive is an important and well-known field of study contributing to the enrichment of science and technology while patent data is an important source of information for company's policymakers and other stakeholders. In this section, the aforementioned fields are described in detail and some previous studies also presented.

[*] *Corresponding author.*
E-mail address: syamin@utm.my

https://doi.org/10.37934/araset.37.2.94103

1.1 Automotive

This has further encouraged many researchers and inventors to develop new technologies for fulfilling the demand in this field. These new technologies are expected to aid in analysing and solving the problems encountered by the major players locally and internationally. Hence, mathematical models have been introduced to simplify the processes involved by employing tools such as statistics, probability theory, graph theory and differential equations. These mathematical methods help in understanding the nature of problems that cannot be clearly interpreted through observation. For instance, in the field of automotive or manufacturing, the dynamic mathematical model has been developed to help medium-sized companies in their sustainable system [1].

Generally, automotive is a field that has provided enough material through possibly more than thousands of books and research articles. Part of the material has been covered in this work [2-4]. The importance of mathematical models in the breakthrough of science and technology cannot be underestimated. Systems such as mechanical, electrical, biological or even economics can be accurately described using a mathematical model. In combination with the rapid development of computers during the last 30 years, the number of available models within every scientific area has expanded. Today, it is possible to numerically solve complex process models, which was hardly imagined a couple of decades ago. Models can be constructed simply while producing the real process behaviour. The models should be accurate and concise besides being able to reveal everything about the internal cause-effect relationships within the process. Each model is commonly built for a specific task to a prescribed accuracy [5].

Unfortunately, mathematically modelling a problem is not as easy as it sounds. Nonetheless, the perspective of the model can serve as a good long-term goal for those who deal with it. In industries that produce large and heavy products like automotive, it must be realised that an accurate model is very important as it requires large and heavy machinery and facilities and involves complex production processes. It should be noted that a mathematical model is a simplification of reality, especially when modelling complex systems involving current trend technology [6]. Usually, theoretical assumptions are required to start a model. An accurate model of a process would allow the prediction of process behaviour for different conditions, thereby enabling the optimisation and control of a process for specific purposes such as risk and profit.

There are several steps to be considered in using mathematical modelling approaches to a problem. Continuous measured data free from noise is very helpful in the validation procedure, an important part of general modelling methodology that considers the nature of a system or process that occurs [7].

1.2 Patent

A patent is a legal document that grants exclusive use and rights to the inventor or assignee. A patent records information about the technology and describes the processes related to the invention [8]. Analysis of patent data could essentially enrich the scope and depth of strategic technology policy-making: alignment of a company’s innovation strategies, evaluation of R&D proposals, assessment of technology competitiveness and other important tasks. The most comprehensive study based on patent data is related to patent landscape.

A patent landscape contains various information regarding trends, geography, patent strategies, leading companies and legal events. Modern patent analytics provides a systemic view for better policy making especially in the context of extreme growth of the world’s repository of patent data and evolution of the intellectual property analytics [9]. Patent analytics (intellectual property analytics) has been widely used to measure innovation performance [10-12], find knowledge spillovers [13,14] and monitor emerging technologies and technology diffusion [15].

2. Literature Review

A clustering algorithm with non-exhaustive overlaps is proposed to overcome deficiencies with exhaustive clustering methods used in patent mining and technology discovery. The non-exhaustive clustering approach allows for the clustering of patent documents with overlapping technical findings and claims, a feature that enables the grouping of patents to define related key innovations [16]. Clustering is an effective text mining technique and is also considered an unsupervised technique with the goal of putting similar objects in groups whose members have more similarities with each other [17]. It is obvious that patent classification and Derwent World Patents Index (DWPI) data are the two most important types of patent information that may greatly improve patent clustering, whereas claims may be an optional type of patent information [18].

Clustering is a continuous process that includes collecting data, determining a similarity criterion between data, selecting an appropriate clustering method, evaluating the performance of the selective method and finally interpreting the results of clustering. In the previous study, K-Means Self-Organising Map (SOM) was adopted to conduct the task of patent clustering analysis [19]. Clustering algorithms are widely used in unsupervised learning and are among the most useful techniques to discover groups of similar items and uncover patterns in the underlying data. In a typical clustering algorithm, data is partitioned and distributed between clusters based on the similarity between groups and dissimilarity among different clusters. Such algorithms work on calculating distance amongst the points, iteratively calculating revised distance among different groups and classifying new data points into identified clusters [20].

Among various machine learning techniques, document clustering is very useful for patent analysis. As a way to improve efficiency including patent analytic work, many studies have applied machine learning technology to patent documents for revealing business trends and performing technical analysis, which has been ongoing for many years [21]. Through clustering, similar patents are grouped and utilised for invalid patent searches, patent summarisation, technology monitoring, technology prediction and the development of patent map [22]. Additionally, various clustering methods have been utilised for patent clustering. Association rule mining (ARM), reinforcement-learning-based page ranking algorithms, support vector clustering, matrix factorisation and the common k-means clustering method have all been used for patent clustering [23-25].

3. Methodology

Normally, the aspects of patent documents are useful for machine learning techniques such as international patent codes (IPCs), patent citation information and structured data regarding inventors including regular text data.

3.1 Data Source

The patent databases were collected from the DWPI database in March 2022. All automotive businesses that filed for a patent in Japan were included in the data collection with the application year starting from 2000 until 2021. The patents were then split into two time periods: 2000-2010 and 2011-2021. This study focused on two automotive companies, which are Mazda and Mitsubishi, totalling 25,595 patent applications.

The primary reason for selecting these two Japanese automotive companies is because their sales ratios for their respective product segments are relatively similar, with an average of 60% going to SUVs and the remaining 40% to other product segments such as compact cars, sedans and MPVs [26,27].

3.2 Text Mining

The DWPI abstract was employed in this text mining to examine the patent trends of patent publications. The text mining process as presented in Figure 1 consists of two main steps:

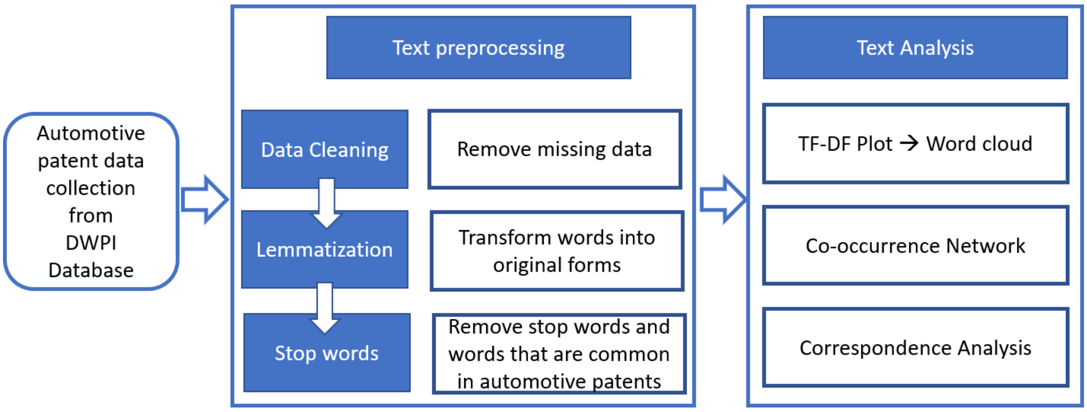


Fig. 1. Flowchart of text mining process

3.2.1 Text pre-processing

There were 244 null abstracts in the dataset. After omitting the missing abstract, a total of 25,351 patents were used for analysis. Numerous abstracts have sentences that are not significant (e.g., "Drawing includes non-English language text"), which would not accurately reflect the key trends from these two companies if included. The sentences were therefore removed from the abstract.

Sentences were broken up into words for text pre-processing. The "Standard POS Tagger" was used for lemmatisation, which transformed these words into their original forms. For instance, the word "include" was retrieved from a text that contains the phrases "include", "includes" and "included". KH Coder provided two forms of word processing, which are stemming and lemmatisation. Stemming cuts off the end of words, while lemmatisation does not treat all words equally. Lemmatisation, for instance, tries to differentiate between the verb "think" and its gerund form, "thinking." This processing lessens the workload needed for counting frequency, figuring out how words relate to one another and generating coding rules.

After that, the stop words were removed. ‘A’, ‘and’, ‘the’, as well as words that are overly prevalent in automotive patents, such as ‘vehicle’, were included in the list of stop words.

3.2.2 Text analysis

TF-DF plot

The extraction algorithm is used to find the key terms that can differentiate between the two periods of patent application, which can be accomplished by comparing the correlation between TF, the total number of times each phrase appears in the data and DF, the total number of documents where each term appears.

This command generates a figure with the TFs represented on the x-axis and the DFs on the y-axis. In this plot, the correlation between the two variables (TF-DF) is typically strong. To provide accurate and meaningful data from the plot, a word cloud was generated using R software. Word clouds are used to visually represent words in the patent application that appear most frequently.

Co-occurrence network

Co-occurrence network is a common technique used in quantitative analysis. The strength of co-occurrence, which is drawn as network edges, is calculated using the Jaccard coefficient as follows

$$C_{Jaccard} = \frac{a}{a+b+c} \tag{1}$$

where a is the number of co-occurrences, b is the number of locations where taxonomy 1 occurs and taxonomy 2 does not, and c is the number of locations where taxonomy 2 occurs and taxonomy 1 does not. The Fruchterman-Reingold approach is used to locate the nodes (words), which aids in creating undirected networks with straight edges [15]. Then, the next step is the detection of edge and centrality of nodes [28].

Correspondence analysis (CA)

Correspondence analysis is one of the principal component analysis techniques that is aimed primarily at categorical data. Principal component analysis techniques are used to reduce the dimensionality, increase interpretability and at the same time minimise information loss. The main objective of CA is to visualise rows and columns of a data table in two- dimensional space in the form of a map [29]. The distance and relative positions of points have a specific interpretation that helps to explore the types of words that have a similar appearance pattern for each variable in this text analysis.

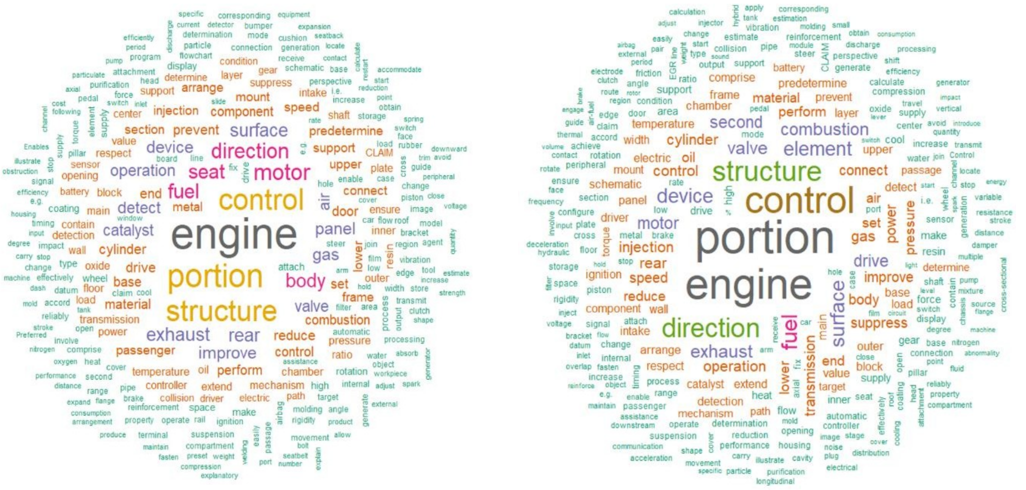
4. Results and Discussion

This section analyses the computational result obtained from the text analysis using KH Coder 3 and R software.

4.1 Key Topics

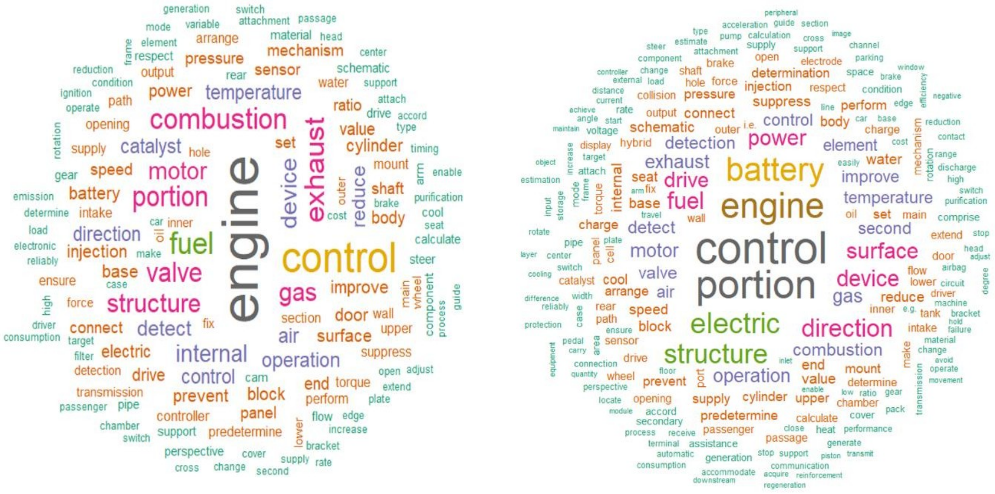
Figure 2 and 3 visualise the frequent words retrieved from abstracts using TF-DF plot from KH Coder and implemented on R software. The least frequent words are represented by the word clouds on the outermost circle (green words). The largest or innermost word, meanwhile, emphasises the most used word in the abstract. The word "engine" was the one that appeared the most frequently in the first period (Figure 2), followed by the words "portion," "control" and "structure." This indicates that Mazda is more engine-focused.

However, its focus has changed from engine only to engine and portion over the past 11 years. Meanwhile, Mitsubishi's primary attention has shifted from engines to controls and portions. Additionally, the battery, engine, electricity and structure are supporting elements in the second term.



Year 2000-2010 Year 2011-2021

Fig. 2. The research topics for Mazda for 20 years starting from 2000

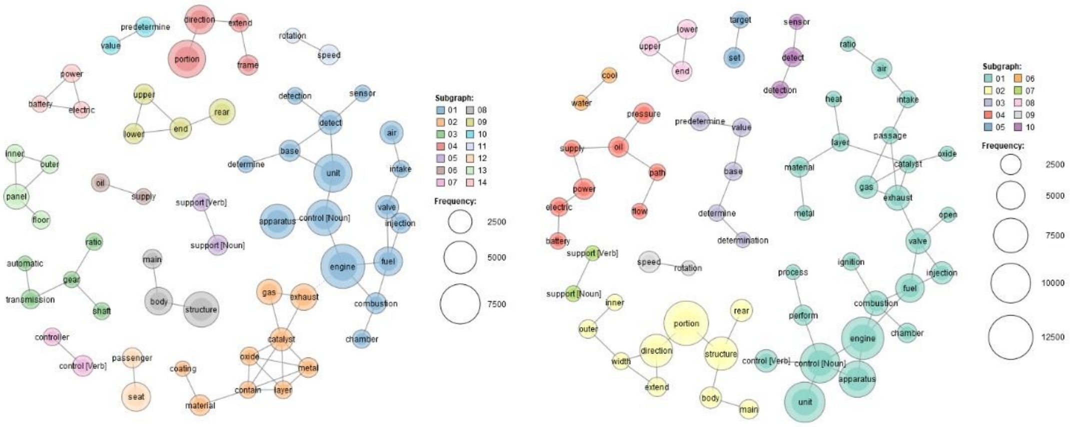


Year 2000-2010 Year 2011-2021

Fig. 3. The research topics for Mitsubishi for 20 years starting from 2000

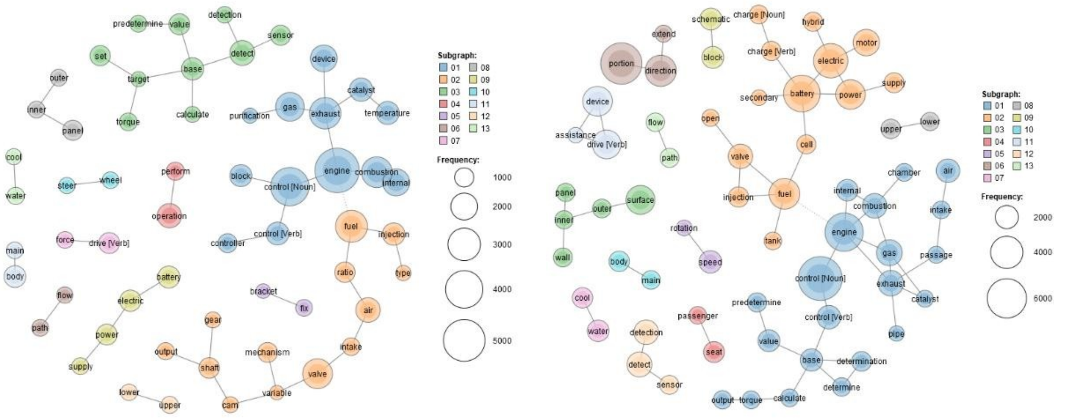
4.2 Co-Occurrence Network

The co-occurrence network commonly occurring patents for Mazda and Mitsubishi is highlighted in Figure 4 and 5. By merely looking at the groups of often occurring words that are frequently combined, the main themes of the data can be identified. The network, often referred to as the connecting line, connects words that are frequently used in combination. If the words are joined by a line, the co-occurrence is strong. The regularity of the qualities is shown by larger nodes. Even if the frequency of the attributes "engine" and "portion" in Mazda was high in Figure 4, there was no co-occurrence between these two terms in the research, suggesting that the attributes "portion" in the red cluster and "engine" in the blue cluster are unrelated.



Year 2000-2010 Year 2011-2021

Fig. 4. Co-occurrence network of frequently occurring patent for Mazda for 20 years starting from 2000

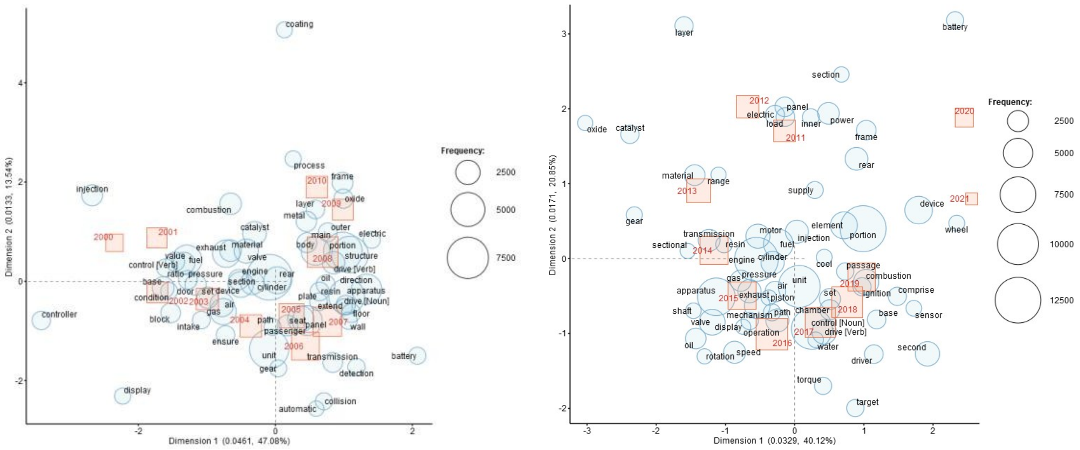


Year 2000-2010 Year 2011-2021

Fig. 5. Co-occurrence network of frequently occurring patent for Mitsubishi for 20 years starting from 2000

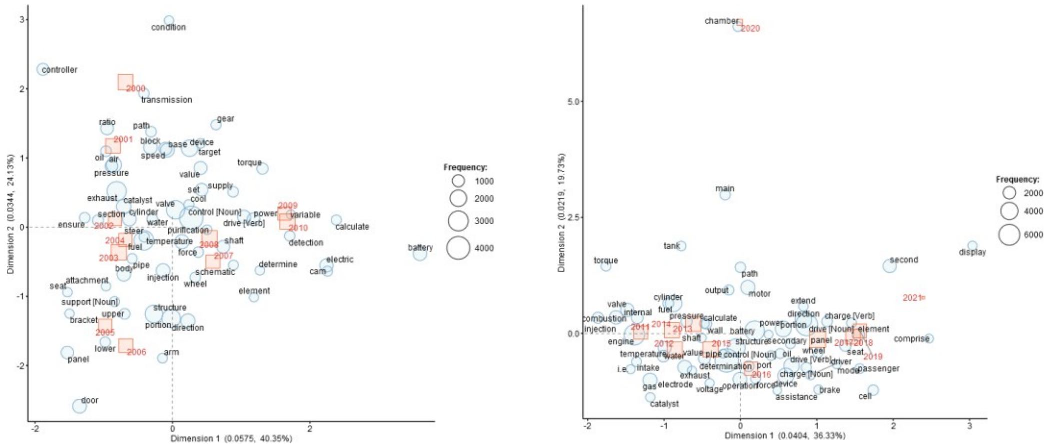
4.3 Correspondence Analysis

The correspondence analysis for Mazda and Mitsubishi, which takes the application year into account, is highlighted in Figure 6 and 7. The two-dimensional relationship between the application year and the words on the same diagram was the focus of the analysis. The area of each circle was proportional to the number of occurrences of each word. As a result, the circle grew in size the more frequently the attribute appears. The number of attributes in the equivalent quantity of text was proportionately represented by the area of each square. In correspondence analysis, uncharacteristic words were uniformly found near the origin, whereas words with strong characteristics were located away from the origin. "Coating" and "battery" were plotted distant from the origin in the first period, indicating their strong features (Figure 6). The second period, however, also demonstrated that the "battery" has strong characteristics. In other words, certain attributes are distinct from others whose contents are different.



Year 2000-2010 Year 2011-2021

Fig. 6. Correspondence analysis for Mazda for 20 years starting from 2000



Year 2000-2010 Year 2011-2021

Fig. 7. Correspondence analysis for Mitsubishi for 20 years starting from 2000

5. Conclusions

This study visualised the frequent words retrieved from abstracts using a TF-DF plot from KH Coder and implemented on R software. The least frequent words are represented by the word clouds on the outermost circle (green words). The largest or innermost word, on the other hand, emphasises the most used word in the abstract. For Mazda, the word "engine" was the one that appeared the most frequently in the first period (2000-2010), followed by the words "portion," "control" and "structure." This indicates that Mazda is more engine-focused. However, the focus changed from engine only to engine and portion for the past 11 years (2011-2021). Mitsubishi's primary attention has shifted from engines to controls and portions in the meantime. Additionally, the battery, engine, electricity and structure were supporting elements in the second term. By merely looking at the groups of often occurring words that are frequently combined, the main themes of the data can be identified. The network, often referred to as the connecting line, connects words that are frequently used in combination. The regularity of the qualities is shown by larger nodes. The correspondence analysis for Mazda and Mitsubishi, which takes the application year into account, has been highlighted. The two-dimensional relationship between the application year variable and the words on the same diagram is the focus of the analysis. The area of each circle is proportional to the number of occurrences of each word. As a result, the circle grew in size the more frequently the attribute appears. The number of attributes in the equivalent quantity of text was proportionately represented by the area of each square. In correspondence analysis, uncharacteristic words were uniformly found near the origin, whereas words with strong characteristics were located away from the origin.

Acknowledgement

This research was supported by the Japan International Cooperation Agency (JICA) through the research grant R.K130000.7343.4B717 entitled ''How Company Manages Intellectual Assets in Open Innovation Era?''. The authors also wish to thank the Intellectual Property and Innovation Management (IPIM) iKohza for the pleasant research atmosphere offered.

References

[1] Thirupathi, R. M., S. Vinodh, and S. Dhanasekaran. "Application of system dynamics modelling for a sustainable manufacturing system of an Indian automotive component manufacturing organisation: A case study." *Clean Technologies and Environmental Policy* 21 (2019): 1055-1071. https://doi.org/10.1007/s10098-019-01692-2

[2] Francis, Abutu, Idris AM, Mohammed Abdulkadir, and Rufai Audu. "THE AUTOMOTIVE INDUSTRY IN NIGERIA: TRENDS, CHALLENGES AND PROSPECTS IN THE 21ST CENTURY." (2017).

[3] Fraga-Lamas, Paula, and Tiago M. Fernández-Caramés. "A review on blockchain technologies for an advanced and cyber-resilient automotive industry." *IEEE access* 7 (2019): 17578-17598. https://doi.org/10.1109/ACCESS.2019.2895302

[4] Rahim, Md Abdur, Md Arafatur Rahman, Md Mustafizur Rahman, A. Taufiq Asyhari, Md Zakirul Alam Bhuiyan, and D. Ramasamy. "Evolution of IoT-enabled connectivity and applications in automotive industry: A review." *Vehicular Communications* 27 (2021): 100285. https://doi.org/10.1016/j.vehcom.2020.100285

[5] Gedam, Vidyadhar V., Rakesh D. Raut, Ana Beatriz Lopes de Sousa Jabbour, Balkrishna E. Narkhede, and Oksana Grebinevych. "Sustainable manufacturing and green human resources: Critical success factors in the automotive sector." *Business Strategy and the Environment* 30, no. 2 (2021): 1296-1313. https://doi.org/10.1002/bse.2685

[6] Yun, JinHyo Joseph, DaeCheol Kim, and Min-Ren Yan. "Open innovation engineering—Preliminary study on new entrance of technology to market." *Electronics* 9, no. 5 (2020): 791. https://doi.org/10.3390/electronics9050791

[7] Sengupta, Abhijit, and Vania Sena. "Impact of open innovation on industries and firms–A dynamic complex systems view." *Technological Forecasting and Social Change* 159 (2020): 120199. https://doi.org/10.1016/j.techfore.2020.120199

[8] Trappey, Charles V., Amy JC Trappey, and Chun-Yi Wu. "Clustering patents using non-exhaustive overlaps." *Journal of Systems Science and Systems Engineering* 19, no. 2 (2010): 162-181. https://doi.org/10.1007/s11518-010-5134-x

[9] Aristodemou, Leonidas, and Frank Tietze. "The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data." *World Patent Information* 55 (2018): 37-51. https://doi.org/10.1016/j.wpi.2018.07.002

[10] Kim, Gabjo, and Jinwoo Bae. "A novel approach to forecast promising technology through patent analysis." *Technological Forecasting and Social Change* 117 (2017): 228-237. https://doi.org/10.1016/j.techfore.2016.11.023

[11] Brem, Alexander, Petra A. Nylund, and Emma L. Hitchen. "Open innovation and intellectual property rights: How do SMEs benefit from patents, industrial designs, trademarks and copyrights?" *Management Decision* 55, no. 6 (2017): 1285-1306. https://doi.org/10.1108/MD-04-2016-0223

[12] Habib, Misbah, Jawad Abbas, and Rahat Noman. "Are human capital, intellectual property rights, and research and development expenditures really important for total factor productivity? An empirical analysis." *International Journal of Social Economics* 46, no. 6 (2019): 756-774. https://doi.org/10.1108/IJSE-09-2018-0472

[13] Hall, Bronwyn H., and Adam B. Jaffe. "Measuring science, technology, and innovation: A review." *Annals of Science and Technology Policy* 2, no. 1 (2018): 1-74. https://doi.org/10.1561/110.00000005

[14] Noailly, Joëlle, and Victoria Shestalova. "Knowledge spillovers from renewable energy technologies: Lessons from patent citations." *Environmental Innovation and Societal Transitions* 22 (2017): 1-14. https://doi.org/10.1016/j.eist.2016.07.004

[15] Fruchterman, Thomas MJ, and Edward M. Reingold. "Graph drawing by force‐directed placement." *Software: Practice and experience* 21, no. 11 (1991): 1129-1164.https://doi.org/10.1002/spe.4380211102

[16] Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. "A brief survey of text mining: Classification, clustering and extraction techniques." *arXiv preprint arXiv:1707.02919* (2017).

[17] Hoo, Chyi-Shiang. "Impacts of patent information on clustering in Derwent Innovation's ThemeScape map." *World Patent Information* 63 (2020): 102001. https://doi.org/10.1016/j.wpi.2020.102001

[18] Bamakan, Seyed Mojtaba Hosseini, Alireza Babaei Bondarti, Parinaz Babaei Bondarti, and Qiang Qu. "Blockchain technology forecasting by patent analytics and text mining." *Blockchain: Research and Applications* 2, no. 2 (2021): 100019. https://doi.org/10.1016/j.bcra.2021.100019

[19] Fredström, Ashkan, Joakim Wincent, David Sjödin, Pejvak Oghazi, and Vinit Parida. "Tracking innovation diffusion: AI analysis of large-scale patent data towards an agenda for further research." *Technological Forecasting and Social Change* 165 (2021): 120524. https://doi.org/10.1016/j.techfore.2020.120524

[20] Park, Hyunseok, Kwangsoo Kim, Sungchul Choi, and Janghyeok Yoon. "A patent intelligence system for strategic technology planning." *Expert Systems with Applications* 40, no. 7 (2013): 2373-2390. https://doi.org/10.1016/j.eswa.2012.10.073

[21] Kim, Mujin, Youngjin Park, and Janghyeok Yoon. "Generating patent development maps for technology monitoring using semantic patent-topic analysis." *Computers & Industrial Engineering* 98 (2016): 289-299. https://doi.org/10.1016/j.cie.2016.06.006

[22] Zhang, Lili, Wenjie Wang, and Yuqing Zhang. "Privacy preserving association rule mining: Taxonomy, techniques, and metrics." *IEEE Access* 7 (2019): 45032-45047. https://doi.org/10.1109/ACCESS.2019.2908452

[23] Beltz, Hayley, Anikó Fülöp, Raoul R. Wadhwa, and Péter Érdi. "From ranking and clustering of evolving networks to patent citation analysis." In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1388-1394. IEEE, 2017. https://doi.org/10.1109/IJCNN.2017.7966015

[24] Jun, Sunghae, Sang-Sung Park, and Dong-Sik Jang. "Document clustering method using dimension reduction and support vector clustering to overcome sparseness." *Expert Systems with Applications* 41, no. 7 (2014): 3204-3212. https://doi.org/10.1016/j.eswa.2013.11.018

[25] Zhang, Lefei, Qian Zhang, Bo Du, Jane You, and Dacheng Tao. "Adaptive manifold regularized matrix factorization for data clustering." In *IJCAI*, pp. 3399-3405. 2017. https://doi.org/10.24963/ijcai.2017/475

[26] Mazda (2022). Company Profile. Retrieved on 14 September 2022. https://www.mazda.com/globalassets/en/assets/about/profile/library/files/2021en_all.pdf

[27] Mitsubishi (2022). Corporate Profile. Retrieved on 14 September 2022. https://www.mitsubishimotors.com/en/company/profile/?intcid2=megadrop_company_profile

[28] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69, no. 2 (2004): 026113. https://doi.org/10.1103/PhysRevE.69.026113

[29] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69, no. 2 (2004): 026113. https://doi.org/10.1103/PhysRevE.69.026113