Journal of Business and Management Studies (JBMS)
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS

Identifying Automotive Industry Trends: Data Mining from Intellectual Property Databases

Roland Attila Csizmazia

Associate Professor, Glocal Education Center, Ingenium College & Department & Kwangwoon University. Seoul, South Korea
✉ **Corresponding Author**: Roland Attila Csizmazia, **E-mail**: csix@gmx.at

ARTICLE INFORMATION

Received: August 18, 2021
Accepted: September 01, 2021
Volume: 3
Issue: 2
DOI: 10.32996/jbms.2021.3.2.12

KEYWORDS

Patent research, trend analysis, text mining, information extraction

ABSTRACT

Patents protect the patent holders in that the patented process, design and invention can only be used or sold by the holder exclusively. Therefore, manufacturers use patents to gain a competitive edge against the competition. A patent analysis discovers which car parts are considered the most for future development in the automotive sector. The purpose of this paper is to identify and analyze major trends and the potential implementation of patents. Furthermore, the research may reveal in detail the common R&D trends within a certain industry, differences among the major representative manufacturers and support to identify feasible future strategies for lagers. The patent analysis will be launched with the data collection from a patent database. To avoid the extensive computing time in R, only each patent document's abstract is deployed for the research. After data cleansing, the term frequency-inverse document frequency algorithm is used to find the keywords in the patent abstracts. To visualize results, the social network analysis is conducted. It identifies trends and relationships among the mapped keywords. The discovered major keywords constitute the graphs of the most important parts related to each other and are considered for the future.

1. Introduction

Text mining has received increasing attention due to the extensive availability of digital textual content. It also facilitates the utilization of patent documents by extracting keywords. The large-scale availability of patent documents may impede the goal of this paper. Accordingly, to remain focused on future movements, either time or the number of documents must be limited. To analyze R, the amount of data will be further decreased by focusing only on the abstract within each patent document. The analysis of patent documents helps find and understand the technology trends and forecast the upcoming trends in research and development. Patent analysis is an objective way to discover trends in technological development (Jun & Lee, 2012). They include information about future development and trends. Once text from patent documents is captured and aggregated, emerging trends are found and explored. Through the trend, analysis networks will be built to visualize and discover the relationships between individual terms within the patent documents for each industry participant. When social networks are represented, the leading development technology forecasting strategies can be discovered. Companies can take advantage of the findings to follow strategies or create their own strategies for development.

2. Literature Review

Patent documents include information on application date, filing date, assignees and inventors, abstract, description of application areas and so forth (Yoon & Park, 2004). Previous researches have already used bibliographic information for technology analysis and forecasting trends: one research dealt with merging technology forecasting using new patent information analysis and the International Patent Classification codes to construct an emerging technology forecasting model (Jun & Lee, 2012); another was conducted to reveal problems by investigating the relationship between patents and to improve patent statistics as a technology indicator (Basberg, 1987); and also to identify characteristics of the patent keyword network and to perform trend analysis to reveal how the role of keywords changes in a network over time (Lee et al., 2015). The importance of keyword selection strategy for text mining revealed which elements of patent documents are to be used for keyword selection, what kind of methods for keyword selection could be used, the number of keywords to be selected and the transformation of keyword selection results into structured data (Noh et al., 2015).

This research attempts to extend the keyword-based morphological patent analysis by a statistic approach through the application of term frequency-inverse document frequency to demonstrate how important a term is to a document within a selected range of corpus. Normalized values enter the document-term matrix to lower-dimensional space. The relationships between terms or keywords will be represented via social network analysis. In such an analysis, structural hole, degree, density, and Euclidean distance are indicators to measure the cohesion of a network, whereas the centrality, centralization, and reciprocity are characteristics of a network (Kim et al., 2014).

3. Methodology

The patent analysis is completed to avoid the use of lengthy computation in R. Merely the abstract of each patent document will serve for the underlying analysis. The analysis is handled through keyword-based analyses, of which the results will be applied in social network analysis to reveal relationships between the selected keywords.

3.1 Data Collection

To visualize the results of research, patent documents in the automotive industry were collected. The target data for the research is available in databases of patent registers, such as the United States Patent and Trademark Office (UPSTO), European Patent Office (EPO), World Intellectual Property Service (WIPSON) in Korea. The data was acquired on WIPSON based on its advanced gratuitous global search services for academic use inside the country. Only the patent documents from the year 2000 were observed, while earlier ones were omitted from the research. The representative automotive manufacturers were selected based on the number of sales in the US markets in 2014.

Company	Raw data	Data
GM	8,530	5,621
Ford	838	700
Toyota	9,232	6,066
Chrysler	566	450
Honda	8,012	5,084

Table 1: Number of applied patent documents of car manufacturers

3.2 From Semi-structured Patent Data to Structured Data

The data from WIPSON (Excel tables) were transformed into comma-separated values for the implementation in R. The tm_map() function takes care of preprocessing of the corpus, i.e., erasing unnecessary white spaces, numbers, punctuations and stop words; transforming text to a single lower case. The words were stemmed to reduce the number of unique items that need to be tracked to speed up computational work.

3.3 Selection of Keywords

The result of stemming are terms that serve the keyword selection process. As classifiers require the data in the form of a table, where each row holds a document (*case*) and each column takes a stemmed term (*attribute*), the document-term matrix is introduced to exploit propositional representation. The keyword selection and restriction is accomplished by applying term frequency-inverse document frequency with the argument of normalizing term frequencies. The weighted matrix provides the top keywords.

3.4 Social Network Analysis

The social network analysis provides a visualization by bipartite networks, which consist of *n* numbers of nodes in rows and *m* numbers of nodes in the column.

4. Results and Discussion

The analysis process is applied in the automotive industry to test data and visualize results for technology development within the industry. Although only abstracts of patent documents were picked for manipulation, the document-term matrix size reached 200 MB and the size of term frequency-inverse document frequency grew up to 900 MB. This led to considerable computation time in R.

4.1 Keywords for each representative company

Based on the term frequency-inverse document frequency algorithm, the document-term matrix for each automotive manufacturer is plotted. The rank of tf-idf is applied to find the top keywords. The lower the rank is, the more relevant is the word for the research. This is carried out for the top five manufacturers. An example is represented by Table 2 for GM.

	word	TFIDF	TF	Rank TFIDF	Rank_TF
2	roof	1.50439	611	52	268
5	oil	1.146862	667	157	248
7	light	1.132861	953	166	185
8	seat	1.109951	2182	186	69
9	hing	1.102593	624	191	264
11	pedal	1.072692	819	216	210
13	rail	1.071337	666	220	250
15	head	1.069617	629	223	262
18	coolant	1.031295	850	276	200

Table 2: Top keywords of GM

4.2 Social Network Analysis Based on Top Keywords

The top keywords served to plot the bipartite networks, which represent the relationship of keywords without any refinement. The edges only describe the existence of a relationship between two keywords irrespective of weighting the relationship. Figure 1 represents such a network graph for GM.

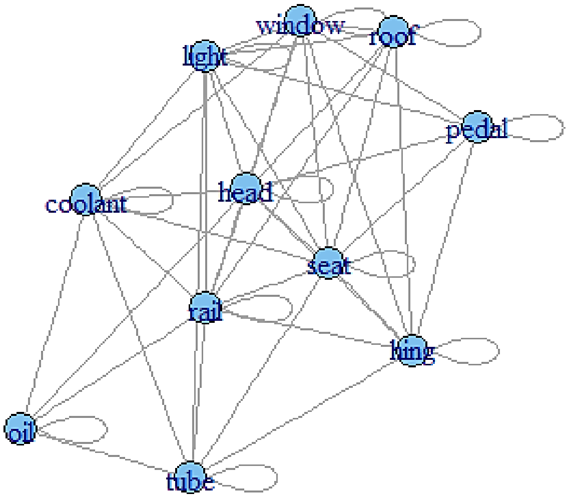

Figure 1: Social network graph of GM's keywords

This graph depicts which parts of cars have been most important. A few of the keywords, e.g., window, seat, roof, light, refer to the increased safety of cars. The relative distance of these keywords to each other shows a strong relationship among them and confirms their affiliation.

The network graph of the keywords extracted from Ford's patent documents is represented in Figure 2. In the case of Ford, the major focus of R&D lies in the improvement of engine performance, fuel consumption and security.

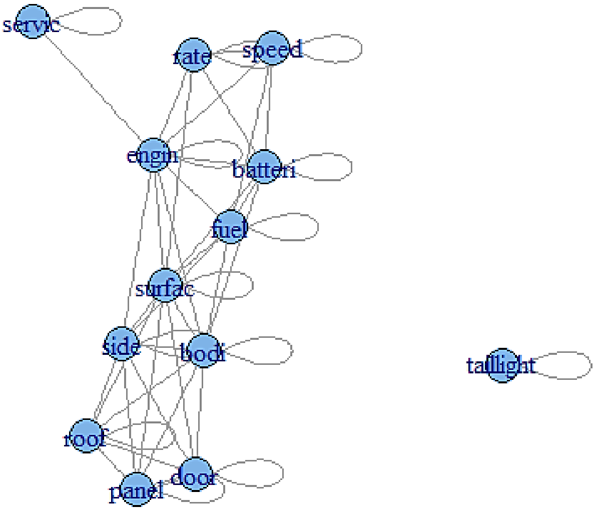

Figure 2: Social network graph of Ford's keywords

Similar to Figure 1, the relative distances among keywords, such as engine, speed, rate, battery and fuel, confirm their strong relationship and may refer to patents to improve engine performance and fuel consumption. Besides, the keywords roof, side, panel, door and so forth are likely to refer to developing the safety of cars.

Toyota's network graph is represented in Figure 3. The relationship of the keywords heat, seat, assist and park may depict the improving comfort of driver and passengers. Moreover, the keywords assist, steer, display, park reveal the development of steering assistance. Similar to US manufacturers, Toyota registered patents to improve security. The relationship among the keywords can explain this: airbag, door, steer, assist and seat.

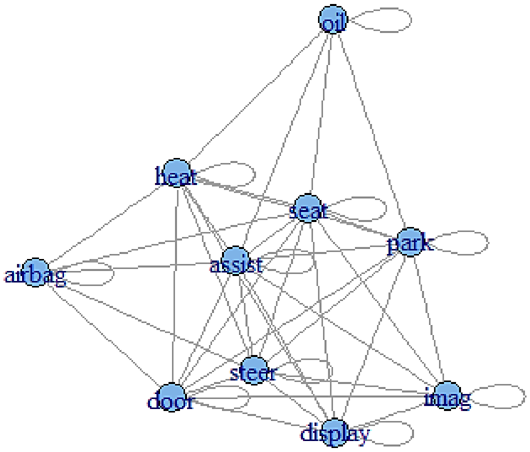

Figure 3: Social network graph of Toyota's keywords

5. Conclusion

This research aimed to discover and identify technology development and R&D trends within the automotive industry. This was achieved based on patent analysis by keyword selection and social network analysis of the selected keywords. The target field was narrowed down to the automotive industry. A further restriction of the patent documents to their abstracts was performed as they hold the analysis information. This reduced the dimensions of the document and the computation time as well.

Nevertheless, the size of datasets became massive to handle by R. The term frequency-inverse document frequency algorithm delivered the keywords for constructing the social networks to analyze the relationship between nodes. Further research work is necessary to find how strongly keywords are associated with each other and how to represent the importance of keywords for each automotive manufacturer.

6. Further Research

Improvement of the research could be achieved by further analysis. Additional research may reveal

- the relative importance of a keyword compared to the other keywords by refining the social network graph based on the degree of node labels, i.e., of keywords;
- the strength of association between keywords could be discovered by applying weighted edges;
- the trends in the entire industry, rather than for each manufacturer within the industry by aggregating datasets for analysis.

Funding: The present research has been conducted by the Research Grant of Kwangwoon University in 2021.

Acknowledgements: This research is based on a joint project originally prepared with Jiho Kang, Zhao Jieying, and Sungjun Park at Korea University.

Conflicts of Interest: "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

References

[1] Basberg, B. L. (1987). Patents and the measurement of technological change: A survey of the literature. *Research Policy, 16*(2–4), 131–141. https://doi.org/10.1016/0048-7333(87)90027-8

[2] Jun, S., & Lee, S.-J. (2012). Emerging Technology Forecasting Using New Patent Information Analysis. *International Journal of Software Engineering and Its Applications*, *6*(3), 107–116.

[3] Kim, J. (1), Choe, D. (1), Kim, G. (1), Jang, D. (1), & Park, S. (2). (2014). *Noise removal using TF-IDF criterion for extracting patent keyword* (Vol. 271). Springer Verlag. https://doi.org/10.1007/978-3-319-05527-5_7

[4] Lee, C., Kang, B., & Shin, J. (2015). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting & Social Change, 90*(Part B), 355–365. https://doi.org/10.1016/j.techfore.2014.05.010

[5] Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications, 42*(9), 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050

[6] Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *Journal of High Technology Management Research, 15*(1), 37–50. https://doi.org/10.1016/j.hitech.2003.09.003