Journal of Global Business Management and Economics (2018) 1(1): 1-10

A Study on Establishing Competitive Advantage Strategies based on Patent Data Investigation using TF-IDF and Network Analysis

Seok Yong Yun (Author)

PhD Student,

Department of IT, Policy and Management

Soongsil University,

South Korea, 06978 Seoul, Korea

icanibe@soongsil.ac.kr

Kyeong Seok Han (Corresponding Author)

Professor of MIS, Soongsil University

Correspondence concerning this article should be addressed to

School of Business Administration

Soongsil University,

South Korea, 06978 Seoul, Korea

Contact: kshan@ssu.ac.kr

Abstract

We need to analyze the data about the competitors to establish strategies to achieve the competitive advantage, but it is very difficult to get the current data about them. However, the patent registration data about the competitors are open to the public while it is protected legally for 20 years. This research paper shows that we can establish strategies practically to achieve the competitive advantage against our competitor based on the big data analysis and machine learning tools. In other words, we showed how to establish strategies against the competitive companies based on the analysis of competitors' technological strategies using the quantitative and qualitative patent data that are open to the public using Frequency Analysis, Arc Analysis, Network Analysis, Heatmap Analysis, TF-IDF, LSTM and so on.

Keywords: competitive advantage, technological strategies, big data analysis, machine learning

Introduction

This research paper explores the value of data that were considered only raw data for information and knowledge and were not fully recognized 10 years age. The Economist(2017), newspaper in England says that the world's most valuable resource is no longer oil, but data. The total amount of data in the world was about two zettabytes in 2011 and about eight zettabytes in 2016, which showed that the amount increased four times in five years (Japanese Ministry of General Affairs, 2012). This could be called as the data explosion in the era of IoT(Internet of Things) and the amount increases geometrically in a short period time.

However, many companies are still using the data analysis as "Descriptive and Diagnostic Analysis" rather than "Predictive and Strategic Analysis" in these days. This research suggests how to establish strategies practically to achieve the competitive advantage based on the big data analysis and machine learning tools (Jun Sunghae(2013)).

We need to analyze the data about the competitors to establish strategies to achieve the competitive advantage, but it is very difficult to get the current data about them. However, the patent registration data about the competitors are open to the public while it is protected legally for 20 years. This research will show how to establish strategies against the competitive companies based on the analysis of competitors' technological strategies using the quantitative and qualitative patent data that are open to the public (Max H. Boisot, Ian C. MacMillan, Kyeong Seok Han (2008).

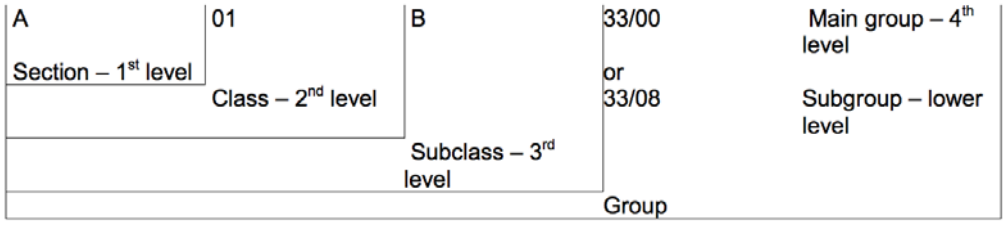
Theoretical Background

Patent Data

Patent systems have the regulations to protect the right of patent holders to promote the invention for the national development. Once the patent passed the evaluation processes, it will be open to the public after one and half years and its right will be exclusively protected for 20 years legally (KIPO(Korean Intellectual Property Office), 2018).

Configuration of Patent Data. Patent data include the quantitative data such as patent number, name, IPC(International Patent Classification), patent applicant, inventor, application date, registration date, open date, international patent application number, international patent application date, abstract, application items, family data, and so on. The patent data also include the qualitative data such as a whole patent description, related diagrams, etc. (KIPRIS(Korean Intellectual Property Rights Information Service, 2018).

Structure of International Patent Classification. International Patent Classification Patent codes were implemented in 1967 with about 70,000 codes and shared by about 180 member countries. As shown in [Table 1] IPC codes consist of Section, Class, Subclass, Main Group and Subgroup. [Table 2] shows that each section includes classified technologies. This scheme is an international standard, even though the system is not perfect.



[Table 1] IPC Code Structure [14]

SECTION	A HUMAN NECESSITIES	B PERFORMING OPERATIONS, TRANSPORTING	C CHEMISTRY, METALLURGY	D TEXTILES, PAPER	E FIXED CONSTRUCTIONS	F MECHANICAL ENGINEERING, LIGHTING, HEATING, WEAPONS, BLASTING	G PHYSICS	H ELECTRICITY
IPC	8,498	16,778	14,449	3,050	3,250	8,551	8,011	8,283

[Table 2] IPC Section Classification

Network Analysis

Network Centrality Theory. The theory is used to identify the importance of a node that is influential in the network. The theory includes three indices i.e. Degree Centrality, Betweenness Centrality, Closeness Centrality (Japanese Ministry of General Affairs(2012). We can identify the strategies and core technologies of the competitors using the network analysis.

Degree Centrality. This is an index to represent the importance level of a node which shows the number of links of the node. If there is a direction edge, we can assess the linkage effect based on In-degree Centrality and Out-degree Centrality. If needed, we can assess the Input and Output effect with the weighted direction edges. The node will be a core node in the network, if the linkage effect value is big enough to be considered as a key node.

Betweenness Centrality. This is an index to represent the number of nodes that are connected through this central node. If a node makes many nodes connected, the Betweenness Centrality of the node becomes very high. It is very important index, because the network will disappear with small Betweenness Centrality.

Closeness Centrality. This is an index to represent the close distance between two nodes. The accessibility and transferability is good, if the distance between two nodes is short as shown in the following equation:

$$C_C(v) = \frac{1}{\sum_{i \neq v} d(v,i)}$$

$d(v,i)$: the distance from node v to node i

Automatic Classification based on TF-IDF

TF-IDF(Term Frequency – Inverse Document Frequency) is a calculation tool to measure the weight for the information search and text mining that can be applicable to the automatic patent classification (Gerard Salton(1988)). TF-IDF suggested by Hans Peter Luhn(1957) is used as a theoretical background for the automatic information classification. The equation is as the following:

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right)$$

$tf_{t,d}$ = frequency of term in documents

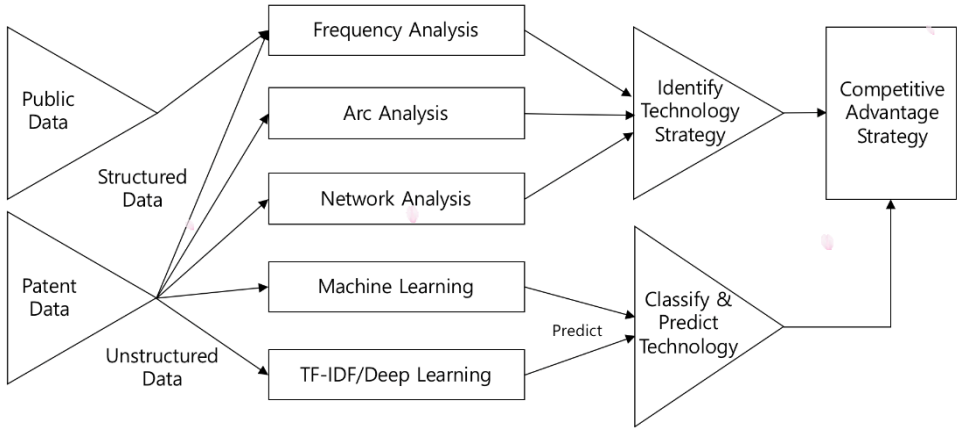
df_t = number of documents containing term

N = total number of documents

$$tf_{t,d} = 0.5 + \frac{0.5 \times f_{t,d}}{\max\{f_{w,d} : w \in d\}}$$

$$idf_{t,d} = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

This research uses 5311 patent data of Company B collected from 2000 to 2014 using KIPRIS(Korean Intellectual Property Rights Information Service). Company B is a competitor of Company A which wants to establish strategies for competitive advantage based on competitor's patent analysis. The data consist of 55 fields such as patent number, name, IPC(International Patent Classification) code, patent applicant, inventor, application date, registration date, open date, international patent application number, international patent application date, abstract, application items, family data, and so on. We collected more data from KISTI(Korea Institute of Science and Technology Information), KIPO(Korean Intellectual Property Office), Google, etc.



[Figure 1] A Research Model to Establish Strategies for Competitive Advantage Based on Competitor's Patent Analysis-PATS(Patent Analytics for Technology Strategy) Model

After collecting the data, we set up a research model to process the big data analysis for establishing strategies to acquire the competitive advantage based on competitor's patent analysis is shown in [Figure 1].

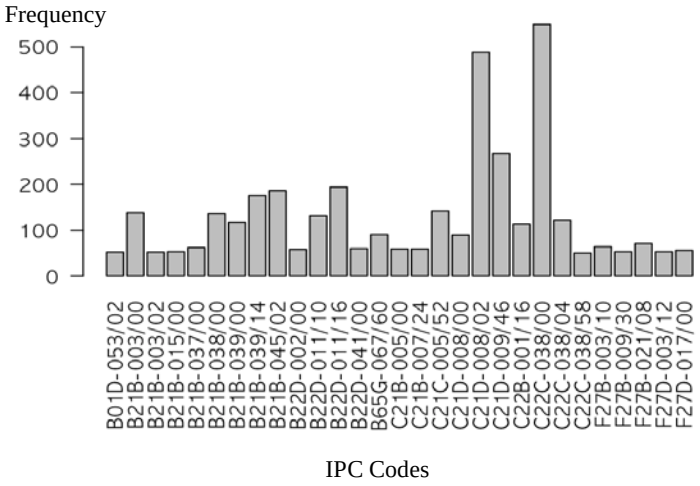
Exploratory Analysis of Competitor's Patent Data

In order to avoid too detailed classification with to small number of patent registrations we perform a frequency analysis of IPC Codes with more than 50 patent registrations.

Patent Technology Analysis

Frequency Analysis

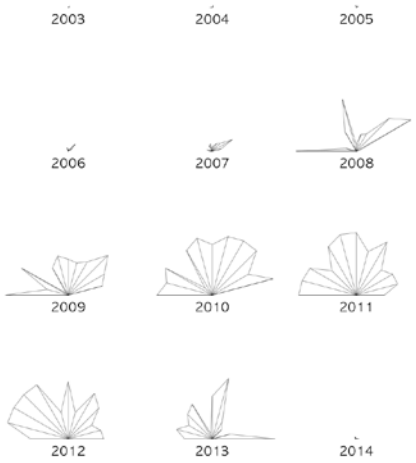
[Figure 2] shows that a few IPC codes such as C22C038/00, C21D008/02 and C21D009/46 have a very high frequency, which means that the competitor is not only developing manufacturing technologies, but also logistics technologies.



[Figure 2] A frequency analysis of IPC codes with more than 50 patent registrations

Plot Analysis

We performed the mosaic plot analysis to find the diversity of patents using yearly-based data. [Figure 3] shows that the numbers and types of patents has increased while the growth of company was realized during the periods.



[Figure 3] Mosaic plots of IPC codes with yearly-based patent registration numbers

Word Cloud Analysis

Word cloud analysis based on the patent names allows us to understand the competitor's technology trends. In the earlier stage the competitor looks interested in 'Rolled Steel' like technologies, but in the later stage it is interested in 'Apparatus' like technologies as shown in [Figure 4].

[Figure 4] The "Word Cloud" based on yearly-based patent registration names

Patent Family Analysis

The patent family analysis provides the insight to understand the competitor's strategies, because a patent family is a set of patents taken in various countries to protect a single invention. In other words, a patent family is the same invention disclosed by a common inventor(s) and patented in more than one country. [Figure 5] shows that the competitor, Company B, registers large patent family in China.

[Figure 5] The geographical distribution of patent registration of the competitor

Patent Network Analysis

The purpose of network analysis is to find out the competitor's key technology.

Centrality Analysis

[Table 3] shows that Closeness Centrality is very low (0.0001), which means that the patents are independent each other in the competitor. However, Degree Centrality is very high (8, 4and 3), which means that the competitor focuses on the steel manufacturing and cold-rolled coil to make automobiles.

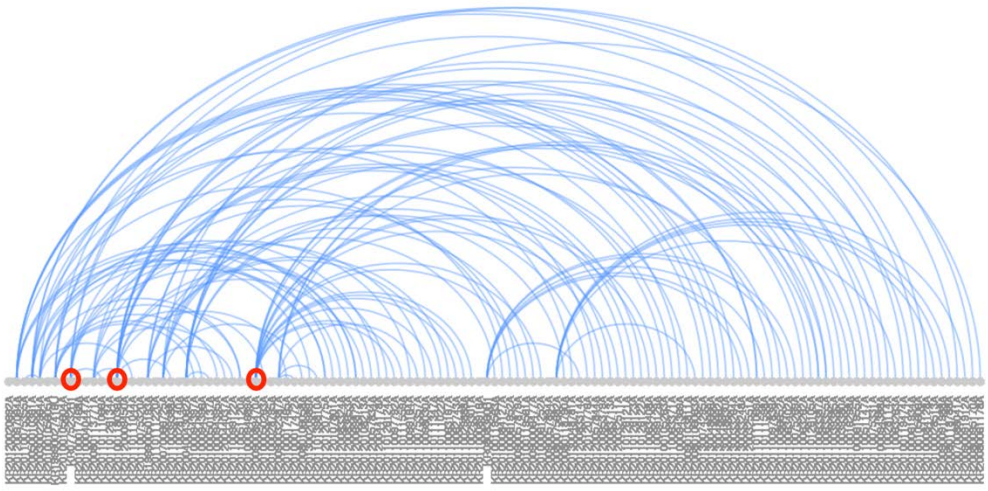
Network Analysis	Statistics	IPC Code
Closeness Centrality	KR2009000***** : 0.0001 KR2010002***** : 0.0001	C23C-022/72 B03C-001/22

Degree Centrality	KR2009002***** : 8 KR2008010***** : 4 KR2012003***** : 3	B22D-011/16 C21D-008/02 B22D-011/124

[Table 3] Centrality Analysis

Arc Analysis

The key patent can be identified through the IPC reference analysis. The competitor, Company B focused on sintering processes and the fuel raw material equipment as shown in [Figure 6]. [Figure 6] also shows that most referred patents are KR2009008***** with 15 times reference and KR2010011***** with 13 times reference.



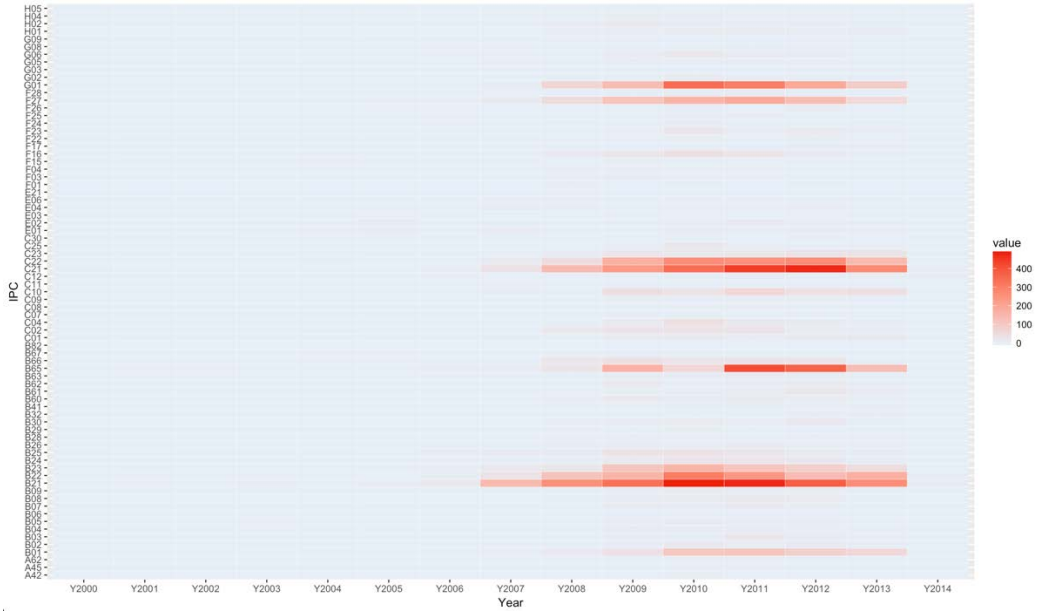
Registered IPCs

[Figure 6] The arc analysis based on registered IPC

Technology Prediction Model Based on the Technology Distribution

Heatmap Analysis

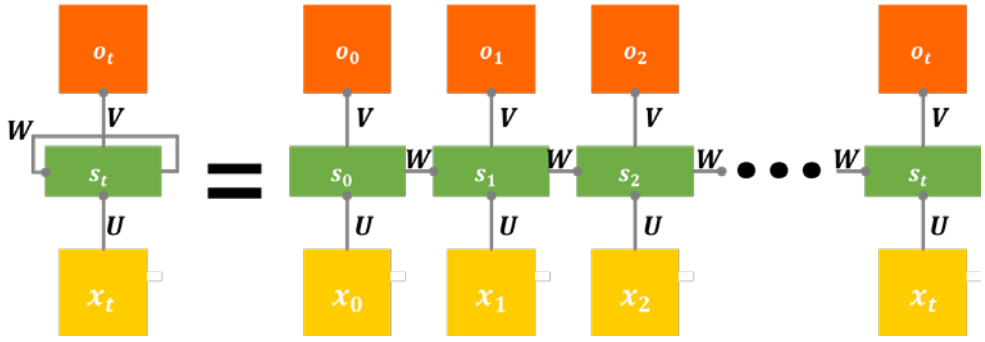
If we draw the frequencies of competitor's patents on each year, we can find that the competitor's patents are developed during certain periods as shown in [Figure 7].



[Figure 7] The yearly based frequency(Heatmap) of competitor's IPC codes

Technology Prediction Based on LSTM(Long Short-Term Memory)

RNN(Recurrent Neural Network) is a method of the deep learning based on neural network. RNN is an applied model evolved from DNN(Deep Neural Network). We want to forecast the competitor's technology strategies based on competitor's patent data and RNN method. [Figure 8] is an example of RNN model.



$$s_t = f(Ux_t + Wh_{t-1}) \qquad o_t = Activation(Vh_t)$$

where, x_t are input values at time t

s_t are hidden layer statuses at time t

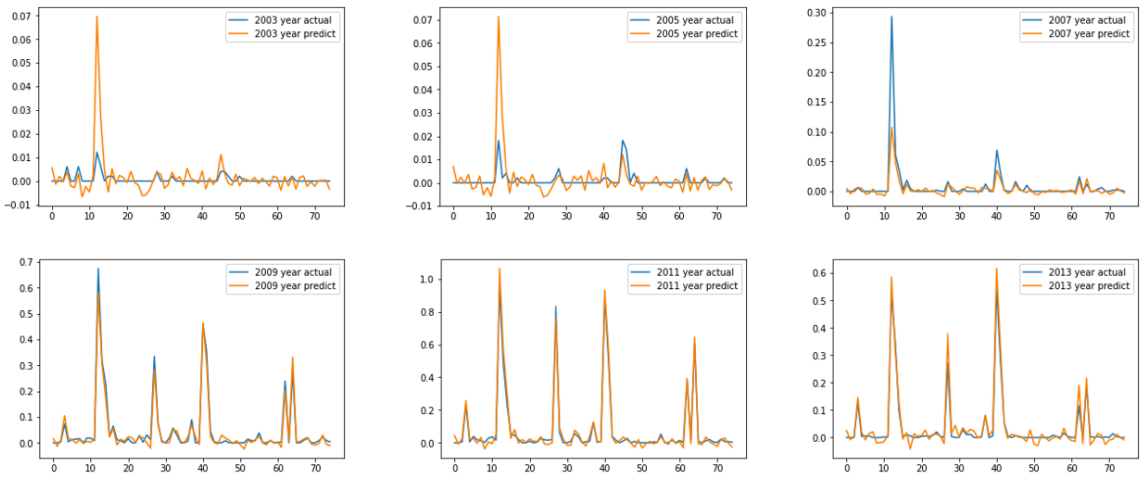
o_t are output values at time t

U, V, W are weight values attained after machine learning

[Figure 8] An example of RNN model

However, RNN has a vanishing gradient problem and we used LSTM(Long Short-Term Memory) model that is an improved RNN model. The forecasting results per year are shown in [Figure 9]. In the graph x-axis represents IPC codes and y-axis represents IPC registration frequencies. The accuracy is 81.82%, which is very

accurate.



[Figure 9] The forecasting accuracy of LSTM model

Automatic Classification of Competitor's Patent Data Based on TF-IDF

[Table 4] shows the main keywords for the processes summarized based on IPC codes, competitor's information, term-document and so on.

Main Process	Sub Process	Code	Keyword
Iron Making	Sintering	0101	Iron Ore, Subsidiary Raw Material, Sized Ore, Sinter, Limestone
	Coke	0102	Coke, Bituminous Coal, Coal, Powdered Coal, Pulverized Coal
	Subsidiary Products	0103	COG, BOG, Gas, Crude Light Oil, Cement
	Blast Furnace	0104	Pig Iron, Torpedo Car
Steel Making	Refining	0201	Iron, Scrap, Subsidiary Raw Material, Oxygen, Conventional Blowing, Argon, Nitrogen
	Refinement	0202	RH, Degassing, Inclusion, LF, Desulfur, Powder
	Continuous Casting	0203	Turndish, Mold, Cooling, Slab, Bloom, Billet
Rolling	Hot Rolling	0301	Reheating Furnace, Roughing Mill, Roll, Cooler, Winding, Hot Coil
	Cold Rolling	0302	Automobile, Annealing, Galvanize, Plating, Steel Sheet, Directional Properties

[Table 4] Classified processes and the keywords

[Table 5] shows the part of patent classification of the competitor calculated by TF-IDF (Term Frequency – Inverse Document Frequency) statistics. The accuracy of the forecasting was 76%, which will be more improved using SVM(Support Vector Machine) and Word2Vec.

Patent Number	Iron Making				Steel Making			Rolling		Domain Expert Judgment
	Sintering	Coke	Subsidiary Products	Blast Furnace	Refining	Refinement	Continuous Casting	Hot Rolling	Cold Rolling	
KR2012006*****	100	200	0	200	300	1365	100	200	0	Refinement
KR2011006*****	36	26	203	28	100	110	200	300	100	Subsidiary Products
KR2011001*****	0	0	0	0	0	10	987	0	0	Continuous Casting
KR2010010*****	0	0	0	74	0	0	1580	100	0	Continuous Casting
KR2009007*****	0	1007	100	100	0	0	0	0	0	Coke
KR2009010*****	0	106	106	106	10	110	300	728	100	Hot Rolling

[Table 5] The part of patent classification of the competitor calculated by TF-IDF (Term Frequency – Inverse Document Frequency) statistics

Conclusion and Future Research

This research paper shows that we can establish strategies practically to achieve the competitive advantage against our competitor based on the big data analysis and machine learning tools. In other words, we showed how to establish strategies against the competitive companies based on the analysis of competitors' technological strategies using the quantitative and qualitative patent data that are open to the public using Frequency Analysis, Arc Analysis, Network Analysis, Heatmap Analysis, TF-IDF, LSTM and so on. In the future this research model will be able to be applied for other industries.

References

Gerard Salton(1988), "Term-weighting approaches in automatic text retrieval", Information Processing & Management Vol.24 Issue 5

Japanese Ministry of General Affairs(2012), "Information Communication Technology White Paper in 2012"

Jun Sunghae(2013), "A Big Data Learning for Patent Analysis", Journal of Korean Institute of Intelligent System Vol23 Issue 5

KIPO(Korean Intellectual Property Office) (2018), http://www.kipo.go.kr

KIPRIS(Korean Intellectual Property Rights Information Service) (2018), http://www.kipris.or.kr

Luhn, Hans Peter (1957). "A Statistical Approach to Mechanized Encoding and Searching of Literary Information" (PDF). IBM Journal of research and development. IBM. 1 (4): 315

Max H. Boisot, Ian C. MacMillan, Kyeong Seok Han (2008), "Explorations in Information Space", Oxford

The Economist(2017), https://www.economist.com/news/leaders/21721656-data- economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource

WIPO(2017), "Guide to the International Patent Classification"

World Intellectual Property Organization, http://www.wipo.int