International Journal of Software Engineering and Its Applications
Vol. 10, No. 1 (2016), pp. 69-78
http://dx.doi.org/10.14257/ijseia.2016.10.1.08

Patent Big Data Analysis by R Data Language for Technology Management

Sunghae Jun*

Department of Statistics, Cheongju University, 360-764, Korea
**shjun@cju.ac.kr*

Abstract

Massive results from researching and developing technologies have been piling diversely such as patent, paper, or news article. To improve the technological competitiveness, most companies have tried to analyze the results efficiently. The amount of the accumulated results is extremely large, so we have difficult to analyze them. But we have to analyze the big data containing the results of developed technologies. To solve this problem, we use R data language as an approach to technology analysis (TA). TA is to analyze the developed result of target technology using qualitative and quantitative methods, such as statistics and Delphi survey. In this paper, we study on a quantitative approach based on statistics, and analyze patent data by R data language. A patent document contains complete information of the developed technology, because patent system protects the inventor's exclusive right for a limited time period. Therefore we propose a methodology of patent big data by R data language for technology management (TM). To illustrate how to apply our research to real filed, we perform a case study. This study contributes to R&D planning and new product development by TM through TA.

Keywords: *R data language, patent big data, technology analysis, technology management*

1. Introduction

Technology management (TM) is a set of activities which are technology analysis, technological innovation and valuation [1]. The aim of TM is to perform R&D planning efficiently and effectively. So the TM is important to improve competitiveness of a company. Technology analysis (TA) is to analyze the results of researched and developed technologies such as patent and paper [2, 3]. Using the TA results, we can forecast future technology or find relationship between technologies for R&D planning or new product development. There are two major approaches to TA, which are qualitative and quantitative methods. The Delphi is a representative methodology for qualitative TA [4, 5], and a popular approach in quantitative TA is patent analysis [6][7]. The Delphi is depended on the experts' experience, so this is subjective TA approach. In comparison, the patent analysis is to analyze patent documents using statistics and machine learning algorithm, this approach is more objective than the qualitative TA approach [8-11]. In this paper, we focus on patent analysis as an objective and quantitative TA approach. Also we introduce an R data science for more efficient patent data analysis. Our R data science is consisted of R project and data science. R project is free and open software for statistical computing and visualization [12]. Data science is to study data as well as big data including data structure, storage, collecting, and analysis [13, 14]. We analyze patent data using data science methodologies, in addition, use R project as analytical software. Therefore we manage technology efficiently and effectively using the TA results. Next

* Corresponding Author

ISSN: 1738-9984 IJSEIA

section introduces our research backgrounds which are R project software, data science, and management of technology. In Section 3, we propose R data science methodology for efficient and effective TM. A case study to illustrate how our study can be used in practical domain is represented in Section 4. Last section describes conclusions and future works related to the proposed research.

2. R Data Language for Patent Big Data Analysis

R is a data language for statistical computing and visualization [12-13]. For the first time, R was developed on the S project at Bell labs [15]. This is an object oriented programming language and has diverse functions for data analysis, so R is good software for data science. R project is comprised of two modules, which are base and packages. When we install R project from the site of R project firstly, the R base module is set. This R base includes the functions for basic statistics and graphics, such as descriptive statistics, linear model, clustering, and some plots. But, it has the limitation for advanced analyses such as support vector machine, evolutionary computation, fuzzy clustering, *etc*. To solve this problem, R provides the package module. The R package extends the ability of R computing and graphics. For example, to use the functions for support vector machine, in addition, we install the package of 'e1071' to R base [16]. Also the package of 'tm' has many functions of text mining for data preprocessing and text data analysis [17]. This package is very useful for big data analysis, because most big data contain text data. Therefore R is efficient and effective language for data science. Figure 1 shows two modules of R data language.

Basic (R base)	• Visualization – scatter plot, box plot, ... • Descriptive statistics – mean, variance, ... • Statistical model – regression, time series, ... • Data manipulation – importing, exporting, ... • Data structure – vector, matrix, array, list, ...
Advanced (R package)	• E1071 – imputation, sigmoid, SVM, SVR, ... • Cluster – fuzzy clustering, PAM, Silhouette, ... • Tm – text mining, stop words, term doc, ... • sna – SNA graph, degree, gplot, component, • sampling – stratified random sampling, ...

Figure 1. Two Modules of R Data Language

Two modules of R make R to be an efficient data language. This combination is useful to use many methods of data science. In this paper, we focus on combining R system and data science. Data science (DS) is the study of data [13]. DS includes all areas about data which are data collection, transformation, architecture, storage, analysis, visualization, and deployment. So DS is interdisciplinary and needs many skills in every area from social science to engineering, or from mathematics to business. Data transformation and analysis are important issues in DS fields, because much of data are unstructured and not numeric, and we should novel patterns from the data using data analysis. In this paper, we focus on data preprocessing for building structured data and data analysis for finding meaningful patterns in TM field. Figure 2 shows the DS composition.

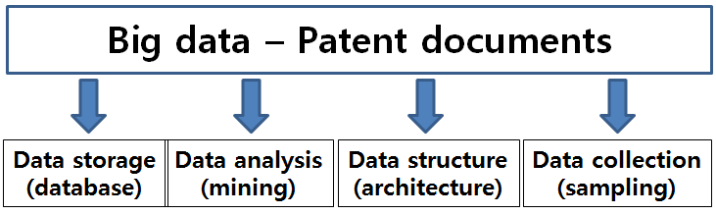

Figure 2. Patent Big Data Science

First, data collection is to gather data from every source such as mobile, social networks, web, document, *etc*. This contains statistical sampling for big data. That is, when we have a difficulty to analyze big data, we should perform sampling from the legacy big data. Second, data storage is to save and manage big data using database system and cloud computing. Next we define and construct data architecture such as data model and data structure. Based on the data structure, we can perform data analysis such as statistics and machine learning algorithm. For example, using the graph of data structure, we make diverse network models such as social network analysis (SNA) [18], and Bayesian network model [19]. In general, we take into account more data analysis than other areas of DS, we use the results of data analysis directly for business and management of technology. In this paper, we apply DS to efficient and effective TM. TM is to manage technological resources of nation and company by interdisciplinary. Generally the technological resources include intellectual property (IP) of patent and paper as well as technology innovation of technological road- mapping and finance. TM is to connect technology and management. Generally technology area deals with the engineering fields such as mechanical technology, bio and chemical, information and communication, medical technology, nano technology, *etc*. The experts of technology pay more careful attention for technology than management. But, they need the mind of management for efficient and effective R&D planning, or technological innovation. In comparison, the experts of management show more concern for management than technology, because the management includes business areas such as finance, account, marketing, strategy, new product development, *etc*. To improve the competitiveness of a company, TM should be performed well. Figure 3 represents the structure of TM.

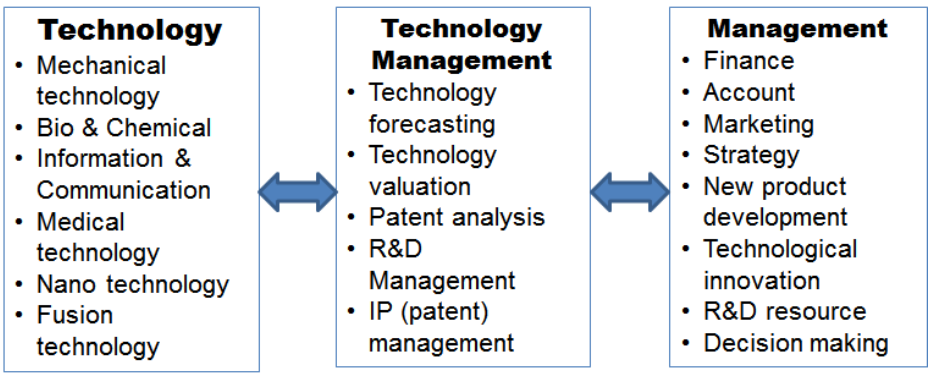

Figure 3. Structure of Technology Management

There were two approaches to TM of a company [1]. First approach was qualitative methodology based on subjective methods such as Delphi. Quantitative methodology by objective methods such as statistical patent analysis was another approach for TM. In this paper, we propose a patent analysis approach for efficient and effective TM. We combine R and DS for patent analysis. Next figure shows the proposed process.

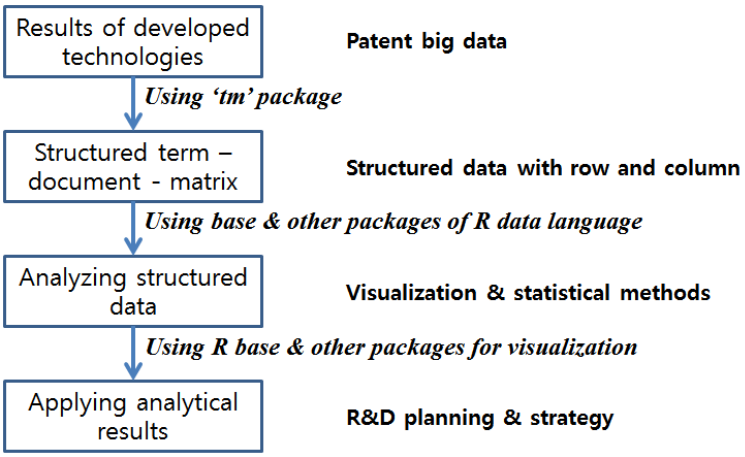

Figure 4. Patent big data analysis by R for R&D strategy

First, we collect technological resources for patent analysis. We mainly search patent document data as technological resource, because a patent document includes detailed results of developed technology such as title, abstract, inventors' name, applied date, claims, drawings, figures, citation, family patents, international patent classification (IPC) codes, *etc*. So a patent document has not numeric data such as text. This is not structured because a structured data includes a table of database (DB) which is consisted of row and column for observation and variable respectively. In addition, each element of table is numeric data. To apply statistics and machine learning algorithm to patent data, we should transform patent document into structured data. In this paper, we use R system for data transformation. The 'tm' package is popular R package for constructing the structured data [17]. Once we get the structured data, we can use diverse analysis methods. The R system includes many functions for data analysis in its base module and packages. First of all, we perform descriptive analysis and draw time series plot. To know the basic characteristic of collected data, these are important. Using the functions provided by R base, we get the results. We use 'sna' package [18] for building advanced statistical model to get the technological relationship between technologies. To know the technological networking is important to TM. We additional use 'xlsx' package to read Excel-type data. Most collected patent data are Excel files. Table 1 shows connection R and data science.

Table 1. R Data Science

Data science	R project	Description & functions
Data manipulation & transformation	tm	Corpus, Document Term Matrix
	xlsx	Excel file import and export
	NLP	Natural language processing
Data analysis & visualization	R base	Mean, variance, visualization,
	sna	Social network analysis, graph
	cluster	Patent and technology clustering

The 'tm' package for data transformation gives some functions such as 'Corpus' and 'DocumentTermMatrix'. For the descriptive statistics, R base module provides 'summary' function for computing mean and variance, and we use the 'sna' package for SNA and graph. Using "cluster" package, we can perform cluster analysis for patent or technology grouping. In addition, the packages of "xlsx" and "nlp" are used for excel file manipulation and natural language processing respectively [21-22]. Also using the result

of patent data analysis, we make efficient and effective R&D planning in the deployment & application phase. We will perform a case study in next section.

3. A Case Study

To verify the performance of proposed methodology, we perform a case study using real technological resource. In this case study, we collected patent documents related patent analysis technology. We collect the patents from the Korea Intellectual Property Rights Information Service (KIPRIS), one of patent databases [23]. The searched patent data were from the Unites States and China. Total number of patents was 86, including 33 the U.S. patents and 53 China patents. Next figure shows the numbers of applied patents of the U.S. and China.

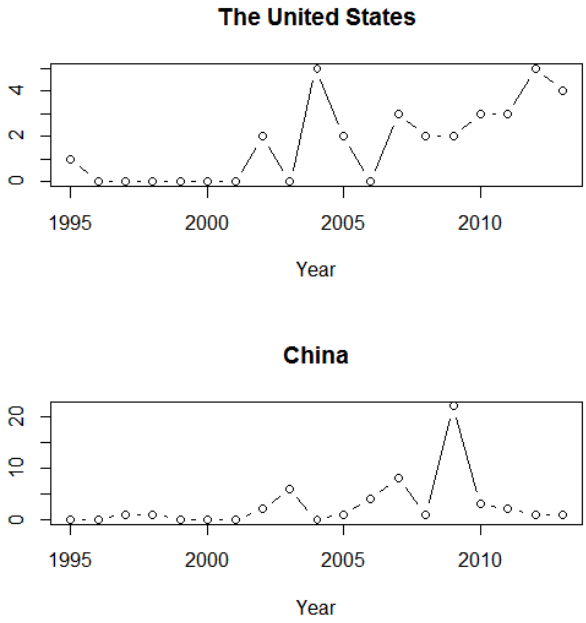

Figure 5. Number of Applied Patents by Year

Both nations developed briskly the technologies related to patent analysis in the 2000s, also applied their developed technologies to patents at the same period of time. Especially, in 2009, the China applied 22 patents related to patent analysis. This was very surprising result. Next we did data transformation and data analysis using the retrieved patent document data. First we should preprocess the collected patent data, because the data were not structured. According to the proposed methodology, we transformed the searched patent data into structured data. Using the 'tm' package of R, we made document term matrix. This consisted of row and column which were patent and term respectively. Each element of the matrix is the occurred frequency of a term in each patent document. Next R codes show the patent data transformation to structured data, patent term matrix.

```
library(tm) # loading tm package
library(xlsx) # loading xlsx package
data=read.xlsx("input.xlsx", 1, header=T)
data.cor=Corpus(VectorSource(data))
data.dtm=DocumentTermMatrix(data.cor)
```

The 'library' function loads R package into R system. The 'xlsx' package can read Excel file to R system by 'read.xlsx' function. The 'Corpus' and 'DocumentTermMatrix'

functions of ‘tm’ package were used for data transformation. Also we extracted keywords from the patent documents. Table 2 shows top 20 keywords of the US and China.

Table 2. Extracted Keywords

Ranking	The US	China
1	Patent	Patent
2	Analysis	Analysis
3	Method	Method
4	System	Module
5	**Claim**	Data
6	**Document**	System
7	Information	Applicant
8	**Search**	Information
9	Display	Database
10	Database	**Classification**
11	Data	**Citation**
12	Module	Family
13	**Present**	Display
14	**Server**	**Comprise**
15	**Report**	**Ranking**
16	**Capturing**	**Computer**
17	Family	**Correspond**
18	**Text**	**Statistics**
19	**Intellectual**	**Client**
20	Applicant	**Software**

The keywords with thicker font are the unique keywords in each nation. The common keyword such as ‘patent’, ‘analysis’, ‘display’, ‘family’, *etc.* represent general technologies for patent analysis. The distinct keywords of the US which are ‘claim’, ‘document’, ‘search’, ‘present’, ‘server’, ‘report’, ‘capturing’, ‘text’, and ‘intellectual’ show the technologies of text analysis and reporting for patent analysis. In comparison, China has the technologies of statistical citation analysis and software for patent analysis by the unique keywords of ‘classification’, ‘citation’, ‘comprise’, ‘ranking’, ‘computer’, ‘correspond’, ‘statistics’, ‘client’, and ‘software’. In this paper, we built more advanced models using SNA. We use ‘gplot’ function of ‘sna’ package to get SNA graph as follow.

```
library(sna) # loading sna package
gplot(keyword.cn.top10, …) # drawing SNA graph
```

Using ‘sna’ package, first of all, we should load this package on R system. The first argument of ‘gplot’ is input data including keywords and next arguments are about the options of SNA graph. Figure 6 shows SNA graph of China using top ten keywords.

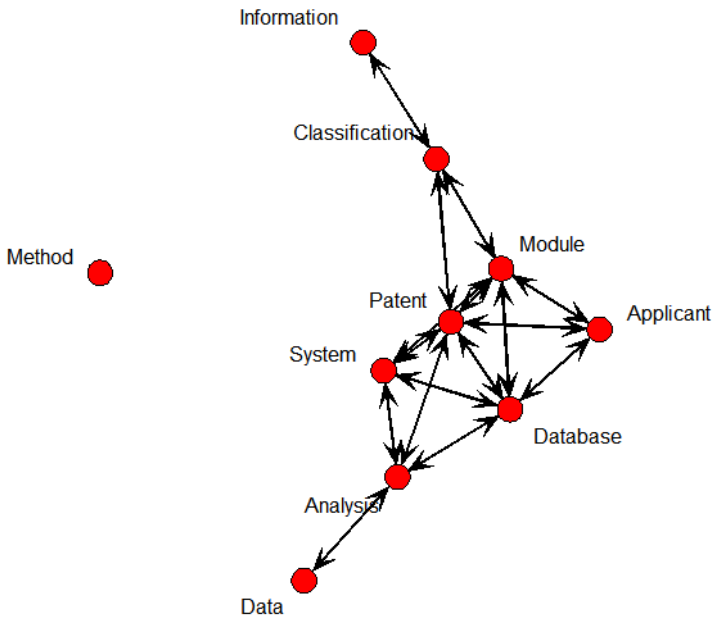

Figure 6. SNA Graph of China using Top Ten Keywords

The SNA graph has two components. Component 1 includes only 'method' keyword, and component 2 includes other all keywords except 'method'. The 'patent' keyword is located in central. There are 'database' and 'system' between 'patent' and 'analysis'. So we knew that the technology of database system is important to patent analysis in China. Next figure shows SNA graph of the US using top ten keywords.

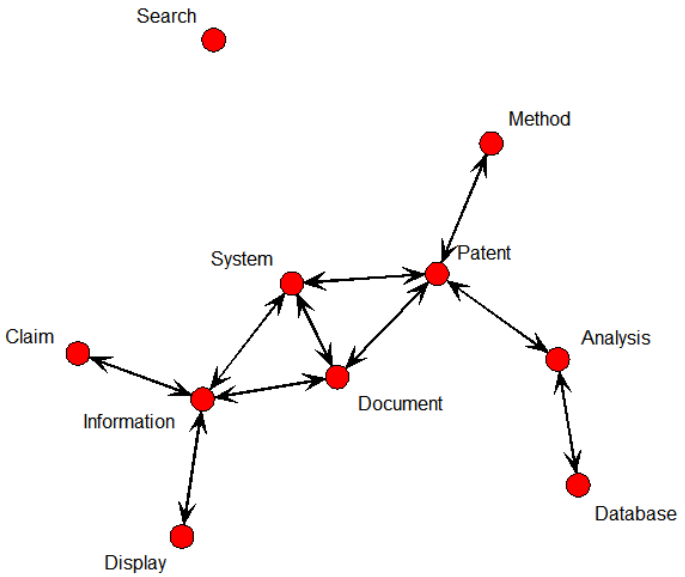

Figure 7. SNA Graph of the US using Top Ten Keywords

Unlike the SNA graph of China, in the SNA plot of the US, the keywords of 'patent' and 'analysis' are connected with each other directly. The 'search' keyword made a component itself. Also the keywords of 'display' and 'information' affect to 'patent' through 'system'. We knew that the 'document', 'system', and method' influence to 'patent' directly. So, we found that the technologies of document system are important to patent analysis. Figure 8 shows SNA graph of China using all keywords.

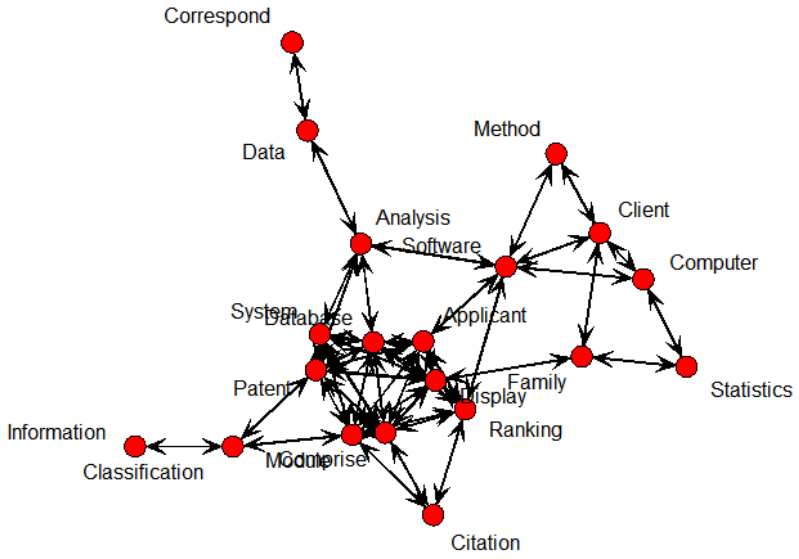

Figure 8. SNA Graph of China using all Keywords

Unlike the SNA graph of China using top ten keywords, this SNA graph has only one component. All keywords are connected with each other. Also the 'patent' keyword is fully connected with many keywords. The 'method' keyword is connected to other keywords through the 'applicant' and 'client' keywords. Figure 9 shows the SNA graph of the US using all keywords.

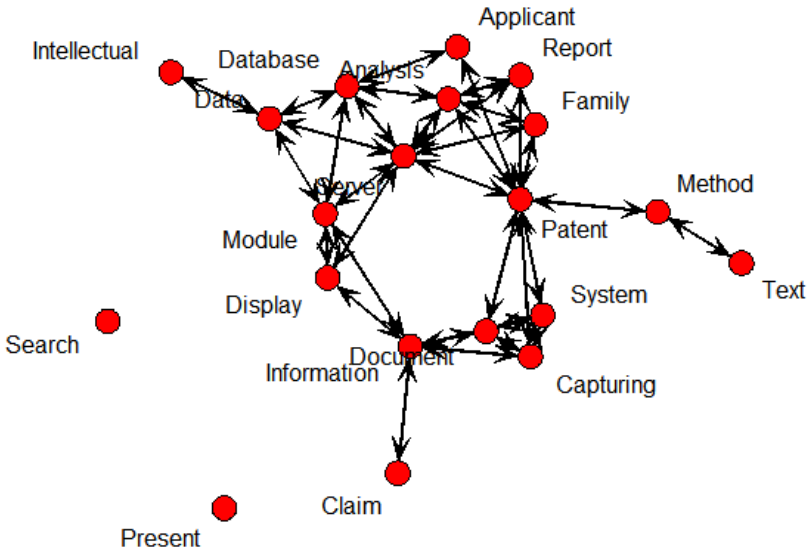

Figure 9. SNA Graph of the US using All Keywords

In this SNA graph, the 'search' keyword constructs a component itself. This result is same to the SNA graph by top ten keywords. Unlike the case of China, the two SNA graphs of the US by top ten and all keywords are similar to each other. Especially the 'server' keyword as well as 'patent' is located in the central. Using the results in this case study, we can build more efficient and effective R&D planning for the technology of patent analysis.

4. Conclusions

In this paper, we proposed a TM methodology using R data science. We combined the R system and data science for efficient and effective TM. The R is opened free software for statistical computing and visualization. In addition, the usage of R package can extend the computational ability of R infinitely. This is strength of R system. Data science is to study about data as well as big data. We used the R as a data language. So, using R data science, we can perform diverse quantitative analyses. In our research, we applied our methodology of R data science to TM. We used the 'tm' package for data transformation, and 'sna' and 'xlsx' packages for patent data analysis. To illustrate how we apply our methodology to real domain, we performed a case study. The target technology was patent analysis, so we collect the patent documents related to patent analysis. We used R data science methodology for this case study. In our future works, we will consider more diverse R packages for patent data transformation and analysis, and apply them to more efficient and effective TM.

Reference

[1] A. T. Roper, S. W. Cunningham, A. L. Porter, T. W. Mason, F. A. Rossini and J. Banks, "Forecasting and Management of Technology", John Wiley & Sons, **(2011)**.
[2] S. Jun, and S. Park, "Examining Technological Innovation of Apple Using Patent Analysis", Industrial Management & Data Systems, vol. 113, iss. 6, **(2013)**, pp. 890-907.
[3] R. Kostoff, D. Toothman, H. Eberhart and J. Humenik, "Text mining using database tomography and bibliometrics: a review", Technological Forecasting and Social Change, vol. 68, **(2001)**, pp. 223-252.
[4] V. W. Mitchell, "Using Delphi to Forecast in New Technology Industries", Marketing Intelligence & Planning, vol. 10, iss. 2, **(1992)**, pp. 4-9.
[5] H. Liimatainen, E. Kallionpää, M. Pöllänen, P. Stenholm, P. Tapio and A. McKinnon, "Decarbonizing road freight in the future – Detailed scenarios of the carbon emissions of Finnish road freight transport in 2030 using a Delphi method approach", Technological Forecasting and Social Change, vol. 81, **(2014)**, pp. 177-191.
[6] Y. Tseng, C. Lin and Y. Lin, "Text mining techniques for patent analysis", Information Processing and Management, vol. 43, no. 5, **(2007)**, pp. 1216-1247.
[7] S. Jun, "Technology forecasting using patent analysis based on cross-impact", Information – An International Interdisciplinary Journal, vol. 16, no. 6(B), **(2013)**, pp. 3853-3864.
[8] S. Jun and S. Lee, "Extracting Key Technology Using Advanced Fuzzy Clustering", International Journal of Software Engineering and Its Applications, vol. 7, no. 4, **(2013)**, pp. 315-322.
[9] S. Jun and S. Lee, "Patent Analysis Using Bayesian Network Models", International Journal of Software Engineering and Its Applications, vol. 7, no. 3, **(2013)**, pp. 205-212.
[10] S. Jun and S, Lee, "Emerging Technology Forecasting Using New Patent Information Analysis", International Journal of Software Engineering and Its Applications, vol. 6, no. 3, **(2012)**, pp. 107-115.
[11] S. Park and S. Jun, "New Technology Management Using Time Series Regression and Clustering", International Journal of Software Engineering and Its Applications, vol. 6, no. 2, **(2012)**, pp. 155-160.
[12] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, **(2015)**.
[13] J. Stanton, "Introduction to Data Science", Syracuse University, **(2013)**.
[14] S. Jun, S. Lee and J. Ryu, "A Divided Regression Analysis for Big Data", International Journal of Software Engineering and Its Applications, vol. 9, no. 5, **(2015)**, pp. 21-32.
[15] Wikipedia, the free encyclopedia, http://en.wikipedia.org, **(2014)**.
[16] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang and C. Lin, "Package e1071", CRAN R Project, **(2015)**.
[17] I. Feinerer, K. Hornik and D. Meyer, "Text mining infrastructure in R", Journal of Statistical Software, vol. 25, no. 5, **(2008)**, pp. 1-54.
[18] C. T. Butts, "Social Network Analysis with sna", Journal of Statistical Software, vol. 24, iss. 6, **(2008)**, pp. 1-51.
[19] S. G. Bottcher and C. Dethlefsen, "Learning Bayesian Networks with R", DSC 2003 Working Papers (Draft Versions), **(2003)**, pp. 1-11.
[20] M. Maechler, and P. Rousseeuw, "Package cluster", CRAN R Project, **(2015)**.
[21] A. A. Dragulescu, "Package xlsx", CRAN R Project, **(2015)**.
[22] K. Hornik, "Package NLP", CRAN R Project, **(2015)**.
[23] KIPRIS, Korea Intellectual Property Rights Information Service, www.kipris.or.kr, **(2015)**.

Author

Sunghae Jun is Professor in the Department of Statistics, Cheongju University, Chungbuk, Korea. He received B.S., M.S., and PhD degrees from Department of Statistics, Inha University, Incheon, Korea in 1993, 1996, and 2001, respectively. He also received PhD degree from Department of Computer Science, Sogang University, Seoul, Korea in 2007, and PhD from Information Management Engineering from Korea University, Seoul, Korea in 2013. He was visiting scholar in Department of Statistics, Oklahoma State University, Stillwater, Oklahoma, USA from 2009 to 2010. His current research interests include big data learning and technology forecasting.