Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/081L05

A Technology Forecasting Method using Text Mining and Visual Apriori Algorithm

*Sunghae Jun**

Department of Statistics, Cheongju University, Chungbuk 360764, Korea

Received: 11 Apr. 2013, Revised: 6 Aug. 2013, Accepted: 7 Aug. 2013
Published online: 1 Apr. 2014

Abstract: Technology forecasting (TF) is the prediction of the future aspect of a technology. TF is therefore an important tool for planning an R&D policy efficiently, and thus most firms and governments consider it to be essential to their technological competitiveness. Since developing technology is usually patented, the efficient analysis of data presented in patent documents is an obvious approach to TF. In this paper, we propose a method for analyzing patent data, using a combination of text mining and the Apriori algorithm. To verify that our method yields an improved performance, we performed an experiment using patent documents concerning database technology retrieved from the United States Patent and Trademark Office.

Keywords: Apriori algorithm, technology forecasting, text mining, patent analysis.

1 Introduction

Technology forecasting (TF) is an approach to predicting the future aspect of a technology [1]. It provides a novel result that can be applied in managing R&D policy. It is, however, difficult to forecast technology. Many results of TF studies have been published [2], most of which used subjective and qualitative approaches, such as Delphi [3, 4, 5]. We definitely need the abundant knowledge of domain experts for TF. However, TF studies conducted by such experts produced inconsistent results because the results were dependent on the experts' experience [6, 7]. To solve this problem, a few research studies, reported in [7, 8], used objective and quantitative TF methods. In [6, 7, 8, 9, 10] data mining techniques and patent documents were used as quantitative methods and objective data, respectively. Data mining is a technique for retrieving novel information from a large database [11] that has been used in diverse fields, such as bioinformatics and customer relationship management (CRM) [11].

A patent is a form of intellectual property (IP). It consists of a set of *exclusive rights* granted by a *sovereign state* to an inventor. It includes complete information of the developed technology. The R&D plan of many firms is based on patent management, that is, the obtaining and maintaining of patents. TF through patent analysis (PA) is an approach to the efficient management of patents [1, 12, 13, 14]. Patent data comprise huge text documents. We can predict future technology of any domain by analyzing the data contained in these documents. However, it is difficult to analyze the documents in their original form using quantitative analysis because, in general, the patent data are neither numeric nor categorical [15]. To overcome this difficulty, in this study we used text mining. In addition, we propose an objective TF method that uses text mining in combination with the Apriori algorithm. In this method, we used our Visual Apriori (VA) algorithm and patent documents as the quantitative method and objective data, respectively. The Apriori algorithm is a popular data mining technique [16, 17, 18]. Our VA algorithm is an extended association mining algorithm based on visualization constructed using extracted association rules. In our previous research, we found that using association rules and maps improved the TF results [19] and used the international patent classification (IPC) codes of patent documents as input data for PA. In the experiment that we performed to verify the performance of our method, we used patent data related to database technology as our given technological domain [19]. In the section 2, we present PA for the purpose of TF, and the Apriori algorithm. Also, we will use keywords of patent documents instead of the IPC codes. We then propose a TF method using text mining and the VA algorithm in section 3. To verify the improved

* Corresponding author e-mail: shjun@cju.ac.kr

Table 1: Lift value explanation

Lift value	Relationship between $Term_x$ and $Term_y$
Greater than 1	Positively associated
0	Independent
Less than 1	Negative associated

performance of our method, we present our experimental results in section 4. The final section presents our conclusion and the direction of future work.

2 Technology forecasting

A patent document includes the complete information about a developed technology, such as the patent number, inventor, international patent classification code, applied date, abstract, title, claims, drawing, citation, and so on. All these details are considered as input data for PA. A popular approach to PA used the analysis of the link structure between patents constituted by citations [20]. However, since it did not analyze the textual information of technology descriptions, this approach was limited in terms of forecasting future technology trends. In this study we therefore selected the title and abstract of patent documents as input data for PA. We forecast future technological trends according to the results of our PA.

In TF, it is very difficult to achieve accurate results, and elaborate methods are therefore required. Our present paper proposes an advanced TF model for efficient technology forecasting using the VA algorithm as a quantitative and objective method. In order to use a VA algorithm, since their original form is not suited to statistical analysis and a machine learning algorithm, we have to transform patent documents into structured data [7,15,21]. In this study, we use the results of our PA method, which uses text mining combined with a VA algorithm, to forecast future technology.

3 Technology forecasting using text mining and visual Apriori algorithm

The VA algorithm comprises association rule mining (ARM) and visualization. Association rule mining is a popular data mining algorithm for extracting novel connections between objects from a large database [11, 16,22]. ARM has two sets of items and transactions. $I = \{i_1, i_2, \cdots, i_n\}$ and $T = \{t_1, t_2, \cdots, t_m\}$ are the items and transaction sets, respectively. A transaction consists of a unique number and contains items [11,13]. A rule of ARM is represented as $(Term_x \rightarrow Term_y)$, where $Term_x$ and $Term_y$ are the objects of transactions. Finally, the extracted rules of ARM are evaluated by support, confidence, and lift measures. The measure of support of objects $Term_x$ and $Term_y$ is

$$P(Term_x \bigcap Term_y) \tag{1}$$

$$\begin{array}{ccccc} & \cdots & term_j & \cdots \\ \vdots & & \vdots & \\ document_i & \cdots & frequency_{ij} & \cdots \\ \vdots & & \vdots & \end{array}$$

Fig. 1: Document-term matrix structure

That is, support is the probability of $Term_x$ and $Term_y$ occurring. The measure of confidence is

$$P(Term_y|Term_x) = \frac{P(Term_x \bigcap Term_y)}{P(Term_x)} \tag{2}$$

This is the conditional probability of Termy given Termx. The last measure of the ARM evaluation is lift:

$$\frac{P(Term_y|Term_x)}{P(Term_y)} = \frac{P(Term_x \bigcap Term_y)}{P(Term_x)P(Term_y))} \tag{3}$$

The lift value is from 0 to ∞, as described in Table 1. In this paper, the results of the VA algorithm are evaluated in the same way as the ARM results. That is, we will use support, confidence, and lift as the measures for evaluating the VA algorithm.

This research study proposes a model which combines a VA algorithm with text mining and multiple regression analysis as an approach to PA for finding the trend of a given technological field; that is, we analyze patent documents in order to achieve an efficient TF result. The input data of our model are patent documents, which consist of text and drawn data. It is difficult to analyze these documents directly using our quantitative method. To solve this problem, we first apply a text mining technique to transform the retrieved patent documents into structured data for use in multiple regression analysis and our VA algorithm.

First, using the determined keyword equation of a given technological field, we retrieve the patent documents related to the domain for which we wish to perform TF. The proposed model uses only the title and abstract from among the diverse information of the retrieved patent documents, which we transform into a document-term matrix for our PA method. This matrix consists of documents and terms as rows and columns, respectively. Each value of the matrix is the frequency of each term in a document. Figure 1 shows the structure of the document-term matrix.

In Figure 1, $frequency_{ij}$ is represented by the number of $term_j$ that occur in $document_i$. In general, the column (term) dimension is much larger than the row (document) dimension. In addition, most values of the frequencies are 0. This matrix is therefore extremely sparse. To overcome this problem, we remove the sparse terms from the matrix. After removing the sparse terms, a revised document-term matrix (rDTM) is obtained. Contrary to the document-term matrix, in the rDTM the dimension of the column is smaller than that of the row. The rDTM has

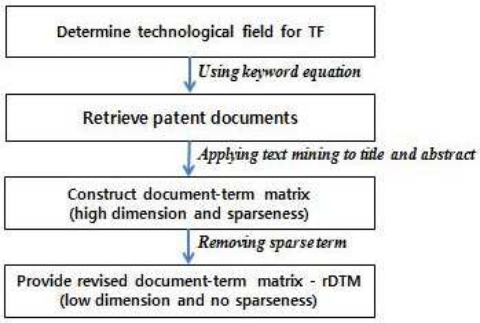

Fig. 2: Process of constructing a revised document-term matrix

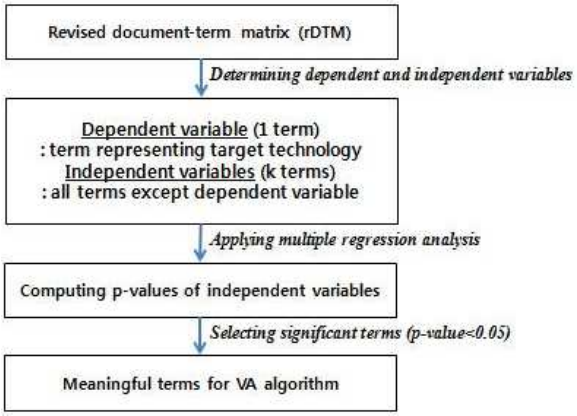

Fig. 3: Process of determining meaningful terms

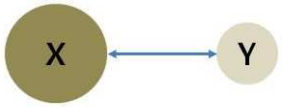

Fig. 4: Visualization of association rule

a low dimension and no sparseness. Figure 2 shows the process of constructing an rDTM.

The resultant rDTM is used for multiple regression analysis and the VA algorithm. We next analyze the rDTM using multiple regression. In a regression model, independent and dependent variables are needed. In this study, the dependent variable is determined as the targeted technological term for the TF. For example, in a TF task that targets database technology, we can determine the term ”database” as the dependent variable. All terms other than the dependent variable (term) in rDTM are considered to be independent variables. These terms are used to explain the technological behavior of the target technology (term). Our regression model is

$$t_{term} = \beta_0 + \beta_1 term_1 + \beta_2 term_2 + \cdots + \beta_k term_k + \varepsilon \quad (4)$$

In this linear equation, t_{term} is the dependent variable, $term_1, term_2, \cdots, term_k$ are independent variables, and and are the regression parameter and error, respectively. The strength between t_{term} (dependent variable) and a term (an independent variable) is represented by a regression parameter. The statistical significance of a regression parameter is interpreted by its probability value (p-value). A regression parameter is significant when its p-value is less than 0.05. Figure 3 shows the process of multiple regression analysis for obtaining meaningful terms.

The results of the regression analysis are used as input for the VA algorithm. In this study, the VA algorithm consists of the Apriori algorithm and the visualization of its results. The Apriori algorithm is an algorithm that extracts association rules by mining frequent object sets [11]. When X and Y are meaningful terms representing technologies, an association $(X \rightarrow Y)$ means that if technology X is developed, technology Y will be developed. This rule is represented as

$$\text{develop technology}(X) \rightarrow \text{develop technology}(Y) \quad (5)$$

We use the three measures of association rules to discover the novel rules from all possible rules. We generate the rules using a support value with a predetermined threshold. The rules are then ranked according to the confidence value. Lastly, we find the final rules by the lift value computed from the results of the support and confidence values. For a more advanced approach to extracting the meaningful rules, we consider the visualization of the result according to the support, confidence, and lift values.

Figure 4 shows a visualization of the Apriori algorithm result. The circle size represents the support value. In addition, the color intensity represents the confidence value. Thus, we can find the novel association rules easily and visually. Combining the Apriori algorithm and visualization, we propose the VA algorithm as follow.

Technology Forecasting using text mining and the VA algorithm

Step1. Constructing revised document-term matrix (rDTM)

(1-1) Determining the technological field for TF;
(1-2) Making the keyword equation;
(1-3) Retrieving patent documents;
(1-4) Using title and abstract of retrieved patent data;
(1-5) Transforming patent data into a document-term matrix;
(1-6) Revising the document-term matrix to obtain an rDTM.

Step2. Selecting meaningful terms using regression analysis

(2-1) Deciding on dependent and independent variables;
(2-2) Modeling regression equation by terms;

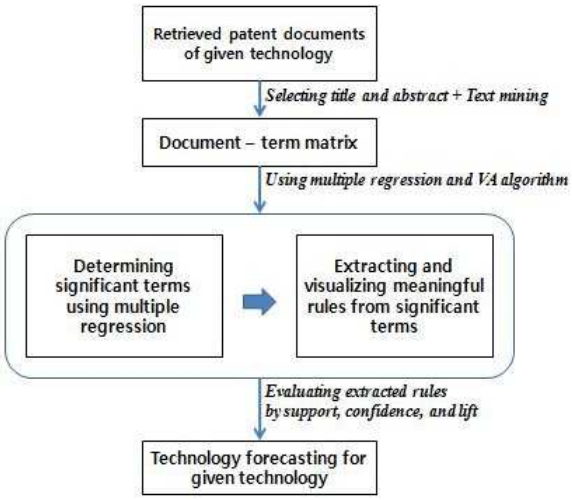

Fig. 5: TF process by text mining, regression, and VA algorithm

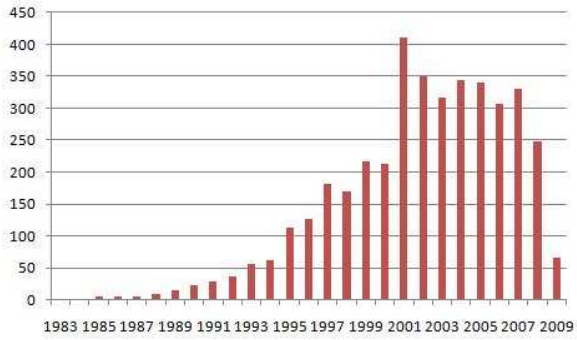

Fig. 6: Number of patents related to database technology

(2-3) Computing p-values of all independent terms;
(2-4) Finding meaningful terms for input of VA algorithm.

Step3. Extracting novel rules for technology forecasting

(3-1) Computing support, confidence, and lift of all rules;
(3-2) Extracting novel rules using Apriori algorithm;
(3-3) Visualizing the result of the Apriori algorithm;
(3-4) Determining final rules for technology forecasting.

In this study, we determine the final rules for TF of a given technology field using the three measures of the Apriori algorithm and the visualization of the results of the Apriori algorithm. Figure 5 shows the complete process of the proposed model.

The process of PA for TF, from retrieving the patent documents to extracting the association rules, comprises three steps: text mining, regression, and the VA algorithm. The experiment we performed to verify the performance of our method is described in the next section.

4 Experiment and result

To verify the improved performance of our method, we used patent documents related to ”database technology” retrieved from the USPTO (United State Patent and Trademark Office, www.uspto.gov) [19]. These data consisted of 3983 patent documents from the beginning of 1983 until July 11, 2011. In this experiment, we used R-project packages for PA [16,23]. Figure 6 shows the number of patent applications filed per year.

The first patent application concerning database technology was filed in 1983. The number of filed patents was increasing in the mid-1990s, and the rate of increase accelerated in the early 2000s. Using all the retrieved patent documents, we constructed transaction data with 206 transactions (rows) and 46 items (columns). We then constructed a document-term matrix. The dimension of this matrix was 3983×16836. That is, the number of documents and terms were 3983 and 16836, respectively. Since most of its values were 0, this matrix was very sparse. To solve this sparseness problem, we reduced the dimension of the document-term matrix. We removed the terms in the bottom 95% of sparseness. Thus, we achieved an rDTM with 3983 documents and 64 terms. The 64 terms are

”*access, accordance, analysis, apparatus, application, associated, automatically, client, communication, computer, control, create, data, database, determine, device, disclose, distributed, executing, file, generate, identifying, information, input, integrated, interface, management, memory, method, model, multiple, network, object, operation, order, performance, plurality, present, processing, program, query, receiving, record, relational, request, response, result, search, selected, server, service, software, specified, storage, structure, system, table, time, transaction, type, update, use, value, and various*.”

We determined the dependent and independent variables from these terms. Since the aim of the TF was to forecast database technology, the term ”database” was selected as a dependent variable of the regression model. All the other terms were used as independent variables. The regression model was defined as

$$database = b_0 + b_1 access + \cdots + b_{63} various \quad (6)$$

where, b_0 is the intercept of the regression model, and $(b_1, b_2, \cdots, b_{63})$ are the parameters representing the strength of the correlation between each independent variable and dependent variable. The p-value of each regression parameter indicates whether the strength is significant. Table 2 shows the meaningful terms and their p-value. All these terms were deemed statistically significant since their p-value was less than 0.05. For TF of database technology, we used these terms as input for the VA algorithm.

Table 2: Selected meaningful terms

Meaningful term	p-value
access	0.000
accordance	0.000
apparatus	0.004
automatically	0.000
disclose	0.000
distributed	0.006
generate	0.009
integrated	0.000
management	0.000
operation	0.027
program	0.006
query	0.000
record	0.008
request	0.000
search	0.010
server	0.000
service	0.029
storage	0.001
structure	0.000
system	0.000
update	0.001
use	0.014
value	0.030

Table 3: Top three rules extracted by support value

Rule	Rank	Support	Confidence	Lift
use → system system → use	1	0.4148	0.7445 0.5648	1.0138
management → system system → management	2	0.3216	0.8834 0.4379	1.2030
storage → system system → storage	3	0.2754	0.7671 0.375	1.0446

Table 4: Top ranked rule by lift and support values

Apriori measure	Extracted rule
Lift=19.4293 Confidence=1 Support=0.0013	{record, request, search, system, use} → disclosure

We then extracted the novel rules using the three measures of the Apriori algorithm. Table 3 shows the top three rules extracted by support value. We found that the three technology (term) pairs, (use → system), (management → system), and (storage → system) were associated. That is, if technology of ”use” was developed, technology of ”system” was also developed. However, the confidence values of ”use” and ”system” were different: the confidence value of rule (use → system) was 0.7445, while, the confidence value of rule (system → use) was 0.5648. Thus, we knew that the technology constructing ”system” was developed after ”use” technology. In the cases of (management, system) and (storage, system), the results were similar to the case of (use, system). Table 4 shows the novel rule with the largest lift value. We found

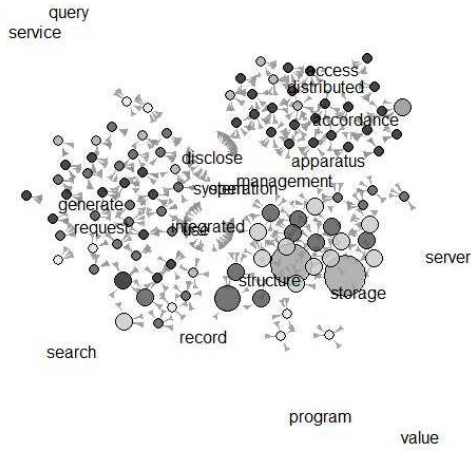

Fig. 7: Apriori visualized result for TF

that the technology of disclosure was developed after the technologies of ”record,” ”request,” ”search,” ”system,” and ”use” were developed.

Figure 7 shows the visualization of the results of the Apriori algorithm. Since the terms ”query,” ”service,” and ”program” are located in the outer reaches of the visualization diagram, it can be seen that their technology was not important. Conversely, since they are in the center of the diagram, it can be seen that the technology of ”system” and ”management” was the basis of all database technologies. Therefore, a company that has this technology will be competitive in the field of database technology. Since, the circle size of ”storage” is larger than that of other terms in the visualization result, ”storage” technology was necessary to database technology. In addition, we can determine that this technology will be needed continuously in the future in the database technology field.

5 Conclusion

In this paper, we proposed a TF method to predict the associated trend of a technology. This study used text mining, regression analysis, and introduced a VA algorithm. In addition, retrieved patent documents were used as input data for the proposed model. In the experiment we performed in order to verify the performance of our method, we selected database technology as the given technology field. We retrieved the patent documents related to database technology from the USPTO. We used only the title and abstract of the patent documents. Using a text mining technique, we constructed a revised document-term matrix, which was used as the input data of the VA algorithm to extract the novel association rules. We also constructed a visualization of the results of the VA algorithm. Finally, we combined the results of the VA algorithm and the

visualization to achieve efficient TF. The results of our research can be applied to any technological field for PA-based TF. In our future work, we will develop a more advanced method for technology forecasting combining diverse data mining and machine learning techniques.

References

[1] V. Coates, M. Farooque, R. Klavans, K. Lapid, H. A. Linstone, C. Pistorius and A. L. Porter, "On the future of technological forecasting," Technological Forecasting and Social Change, **67**, 1 (2001).

[2] A. T. Roper, S. W. Cunningham, A. L. Poter, T. W. Mason, F. A. Rossini and J. Banks, Forecasting and Management of Technology, Wiley, (2011).

[3] V. W. Mitchell, "Using Delphi to Forecast in New Technology Industries," Marketing Intelligence & Planning, **10**, 4 (1992).

[4] Y. C. Yun, G. H. Jeong and S. H. Kim, "A Delphi technology forecasting approach using a semi-Markov concept," Technological Forecasting and Social Change, **40**, 273 (1991).

[5] C. N. Madu, C. H. Kuei and A. N. Madu, "Setting priorities for IT industry in Taiwan-A Delphi study," Long Range Planning, **24**, 105 (1991).

[6] S. Jun, S. Park and D. Jang, "Forecasting Vacant Technology of Patent Analysis System using Self Organizing Map and Matrix Analysis," Journal of the Korea Contents Association, **10**, 462 (2010).

[7] B. Yoon and Y. Park, "Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information," IEEE Transactions on Engineering Management, **54**, 588 (2007).

[8] S. Jun and D. Uhm, "Patent and Statistics, What's the connection?" Communications of the Korea Statistical Society, **17**, 205 (2010).

[9] M. Fattori, G. Pedrazzi and R. Turra, "Text mining applied to patent mapping: a practical business case," World Patent Information, **25**, 335 (2003).

[10] K. Kasravi and M. Risov, "Patent Mining - Discover y of Business Value from Patent Repositories," Proceedings of 40th Annual Hawaii International Conference on System Sciences, 54 (2007).

[11] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, (2001).

[12] D. Zhu and A. L. Porter, "Automated extraction and visualization of information for technological intelligence and forecasting," Technological Forecasting and Social Change, **69**, 495 (2002).

[13] D. L. Mann, "Better technology forecasting using systemic innovation methods," Technological Forecasting and Social Change, **70**, 779 (2003).

[14] J. P. Martino, "Technology forecasting-An overview," Management Science, **26**, 28 (1980).

[15] Y. H. Tseng, C. J. Lin and Y. I. Lin, "Text mining techniques for patent analysis," Information Processing & Management, **43**, 1216 (2007).

[16] M. Hahsler, B. Grun and K. Hornik, "arules-A Computational Environment for Mining Association Rules and Frequent Item Sets," Journal of Statistical Software, **14**, 1 (2005).

[17] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207 (1993).

[18] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast discovery of association rules, In Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, (1995).

[19] S. Jun, "IPC Code Analysis of Patent Documents using Association Rules and Maps-Patent Analysis of Database Technology", Communications in Computer and Information Science, **258**, 21 (2011).

[20] K. V. Indukuri, P. Mirajkar and A. Sureka, "An Algorithm for Classifying Articles and Patent Documents Using Link Structure", Proceedings of International Conference on Web-Age Information Management, 203 (2008).

[21] Y. Tseng, D. Juang, Y. Wang and C. Lin, "Text mining for patent map analysis", Proceedings of IACIS Pacific Conference, 1109 (2005).

[22] M. W. Brinn, J. M. Fleming, F. M. Hannaka, C. B. Thomas and P. A. Beling, "Investigation of forward citation count as a patent analysis method," Proceedings of Systems and Information Engineering Design Symposium, 1 (2003).

[23] R Development Core Team.: R, A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org, (2011).

Sunghae Jun is associate professor in department of Statistics, Cheongju University, Korea. He received the BS, MS, and PhD degrees in department of Statistics, Inha University, Incheon, Korea, in 1993, 1996, and 2001. Also, he got PhD degree in department of computer science, Sogang University, Seoul, Korea in 2007. He has researched statistical learning theory and evolutionary algorithms and is interesting on management of technology (MOT).