

Patent Management for Technology Forecasting: A Case Study of the Bio-Industry

Sunghae Jun

Department of Statistics, Cheongju University, Chungbuk 360764, Korea
and

Sang Sung Park† and Dong Sik Jang

Division of Information Management Engineering, Korea University, Seoul 136701, Korea

Received 29 January 2012, revised 2 April 2012

Patents are an integral part of intellectual property rights and contain substantial amounts of information on developed technologies. Thus, patent management plays an important role in addressing the competitive environment surrounding intellectual property rights. Technology forecasting is an important part of patent management. Any successful R&D strategy requires effective technology forecasting. This paper proposes a technology forecasting method based on various data mining techniques for analysing patents and a patent management approach using the result of technology forecasting. For case study, biotechnology-related patent documents from the US Patent and Trademark Office have been considered. The performance of the proposed method has been verified by this case study.

Keywords: Patent management, technology forecasting, biotechnology, data mining technique

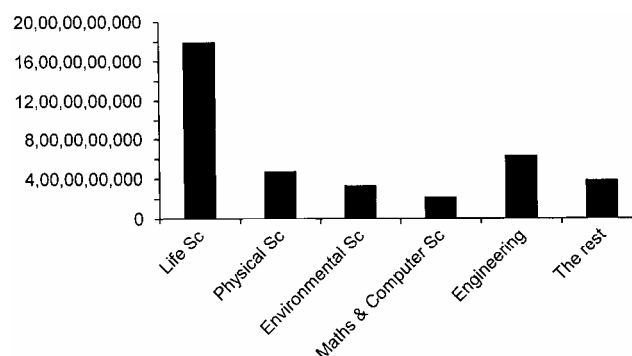
In recent years, the field of biotechnology has expanded as a result of synergies between biotechnology and other fields such as information technology.¹ Bioinformatics is an interdisciplinary field involving biotechnology. Virtually all countries have made efforts to enhance their competitiveness in biotechnology,^{2,3} and most bio-related technologies are either patented or undergoing the patenting process.² In addition, firms generally place great emphasis on protecting their patents. In this regard, a firm's ability to forecast the future of a particular technology plays a critical role in its efforts to enhance its technological competitiveness. This paper proposes a method for technology forecasting (TF), which involves the prediction of the future of a particular technology.⁴ The authors also suggest a patent management approach using the result of TF. Patent analysis (PA) is a novel approach to efficient TF. This research combines three PA methods based on time series analysis, association rule, and patent clustering. So, efficient patent management is possible from the result of TF using PA. In particular, this paper focuses on vacant TF in the bio-industry because the US government earmarked approximately

47 per cent of its total R&D budget in 2000 to life sciences, including biological sciences, agricultural sciences, and medicine (Fig. 1).

Biotechnology has been considered one of the most important technology fields.³ Thus, this research considers patent management in the context of biotechnology. The paper verifies the performance of the proposed approach using retrieved patent documents related to biotechnology from the US Patent and Trademark Office (USPTO).⁵

Objective of the Study

Previous studies included an International Patent Classification (IPC) code analysis for TF in



Source: The National Science Foundation, 2002

Fig. 1 – R&D budget of the US government (2000)

†Corresponding author: Email: hanyul@korea.ac.kr

databases.⁶ This was the first attempt to use association rule mining (ARM) model for TF involving IPC code data. But, this TF model had a limitation because the TF approach was depended on only the ARM model, although, the TF result gave some meaningful rules between the IPC codes. A time series model was also studied for TF of biotechnology.⁷ In this work, the predictive performances of linear regression, Poisson regression, support vector regression, and the moving average (MA) method were compared. The performance of MA was expected to be the best in the competitive methods. Like the result of IPC code analysis, this had a limitation for TF because it also was dependent only on the MA model. To overcome limitations of IPC code analysis based on ARM and time series model based on MA method, a PA method was researched using ARM and Box-Jenkins model.⁸ This was the authors' improved approach to more efficient TF. The authors tried to predict the future technological-trend for patent management such as R&D planning. But, this approach was depended on the experts' experience subjectively because of the lack of objective results based on quantitative (objective) PA methods. So, the TF results could not be connected to patent management effectively. More objective PA approaches and a proper way to connect the TF results in patent management were needed.

Therefore, the current study will develop a more improved TF model for patent management. A PA clustering has been added to previous researches and a way to connect the TF result to patent management sought. This paper proposes a TF method and verifies its performance in terms of its ability to facilitate the development of effective R&D plans. The proposed method can avoid overlaps in R&D investment and facilitate the development of vacant technologies (i.e., undeveloped and incomplete technologies with future potential). For efficient patent management in the context of TF, this paper considers time series analyses, ARM, and patent clustering to construct the proposed TF method for bio-industry. This method is expected to be a useful tool not only for biotechnology but also for other fields.

Retrieved Patent Documents in Biotechnology

In this study, patent documents were retrieved from the USPTO using a search query based on all international patent classification (IPC) codes related to biotechnology. All USPTO patent documents are classified according to IPC codes for each technology

Table 1 – Patent applications in biotechnology

Year	Number of applications
1965-1980	1,136
1981-1985	2,677
1986-1990	4,114
1991-1995	11,387
1996-2000	17,356
2001-2005	12,699
2006-2009	1,517
Total	50,866

field.⁹⁻¹⁰ Biotechnology-related IPC codes (Sections A, B, ..., H) from the World Intellectual Property Organization (WIPO) were selected.¹⁰ A total of 50,866 patent documents were searched for the analysis. The number of biotechnology patents increased until 2000 but then decreased until 2009 (Table 1). There were large numbers of biotechnology projects from the mid-1990s to early 2000 (ref. 3, 11).

As discussed earlier, these increases in the number of biotechnology projects were due to the US government's R&D budget for 2000 (Fig. 1). In this study, to forecast the future the biotechnology, biotechnology-related patent documents retrieved from the USPTO were analysed.

Patent Management and Technology Forecasting

Traditional patent management involves applying, issuing, obtaining, maintaining, and protecting patents.¹² Effective patent management requires strong technical and legal skills. Recently, the goal of patent management has become business-oriented.¹² Patent management allows firms to efficiently plan R&D activity and is a source of competitive advantage for intellectual property rights among competitors like, say, Samsung and Apple. In addition, new R&D paradigms can be derived through patent management,¹³ and R&D policies can be designed accurately and effectively by analyzing existing patent documents.¹⁴ Further, any successful R&D strategy requires careful consideration of future technology trends. Analysing patents can be a shortcut for TF,¹⁵ and patent management is an important topic in technology management. Under technology management, technology marketing, innovation, and new product development depend on patent management.¹⁶ Novel technologies (or patents) are a critical issue in technology management because a firm's patents represent the its technological

competitiveness in this era of patent wars.¹⁷ In this regard, this study proposes a TF method of patent management based on an analysis of biotechnology-related patent documents. For this, the study considers time series analyses of the frequency of applied patents, the ARM of IPC codes,⁴ and the data clustering of abstracts of patent documents, which are combined for biotechnology forecasting. In the next section, a time series model for forecasting the frequency of patent applications is presented.

Time Series Model for Biotechnology Forecasting

For efficient biotechnology forecasting, a time series model was constructed.⁷⁻⁸ For this, the MA and autoregressive integrated moving average (ARIMA) model were considered to predict the future frequency of biotechnology-related patents. The ARIMA model provided good result in TF tasks.⁸ The ARIMA model is a Box-Jenkins time series model¹⁸ and is defined as follows¹⁹:

$$Z_t \sim \text{ARIMA}(p, d, q) \quad \dots(1)$$

where Z and t represent time series data and the current point in time, respectively; d is the number of differencing, p and q are the orders of the autoregressive (AR) and MA parts of the model, respectively. To fit the model, optimal p , d , and q are selected. The ARIMA model requires a stationary times series,¹⁸ that is, no changes in the mean and variance of time series data over time. However, most time series are not stationary and have to be transformed into stationary ones. For this, the first or second differencing of time series data can be used. The Akaike information criterion (AIC) was used to evaluate the fitted ARIMA model.¹⁸ The ARIMA model was determined by the lowest AIC value (training patent data) and evaluated using biotechnology-related patent documents (test patent data). The number of patents related to biotechnology issued per year is illustrated in Fig. 2. The number increased up to the early 2000s. This growth was related to the Human Genome Project (HGP) of the US. The National Institutes of Health began the US HGP and completed it in 2003 (ref. 20). Hence, the trend in biotechnology patent applications depended on the development of HGP.

These time series were non-stationary, that is, there were changes in means over time. They were transformed into stationary time series through first differencing. The time series data from 1972 to 2008

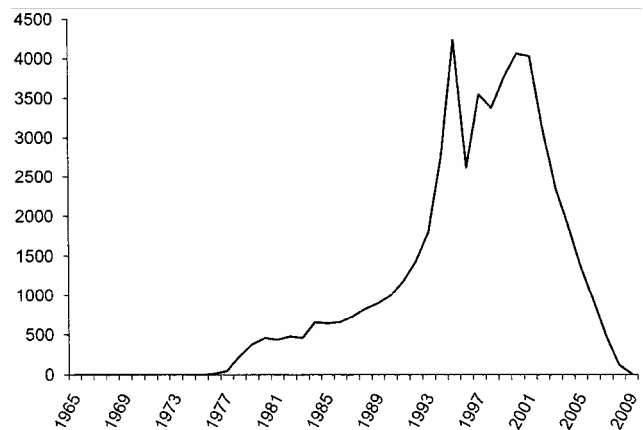


Fig. 2 – Number of patents related to biotechnology

Table 2 – Fitted ARIMA model

(p,d,q)	AIC	S.E.
1,1,1	557.85	537.59
1,1,2	557.28	524.56
2,1,1	557.20	524.06
2,1,2	559.33	532.44

were used to construct the ARIMA model and the optimal ARIMA model $(p,d,q) = (2,1,1)$ selected (Table 2).

The ARIMA (2,1,1) model had the lowest AIC value and the smallest standard error (S.E.), and was thus identified as the optimal model for biotechnology forecasting. The fitted model was then employed to predict the future frequency of biotechnology-related patent applications.

Extraction of Meaningful Association Rules for Biotechnology Forecasting

The ARM was a good PA method to analyse the IPC code data for TF.^{6,8} So, in this study, the ARM was used to extract the meaningful rules from the IPC code data related to biotechnology.

ARM is a popular data-mining tool for extracting association rules and correlated relationships between items.²¹ A data set for ARM consists of items and transactions, which represent variables and observations, respectively. The i -th item (variable) has a binary value $X_i = (0,1)$. For a transaction (observation), if the i -th item occurs, then the value of X_i is 1, whereas $X_i = 0$ if the item does not occur in the transaction. In this study, to extract association rules from a large number of transactions, the following measures were used: support, confidence, and lift. Under the antecedent (A) and consequent (B) items (IPC codes in this paper) of ARM, an

association rule ($A \rightarrow B$) is defined as follows: If technology A is developed, then technology B is also developed. The support measure of ($A \rightarrow B$) is the probability of all transactions containing A and B.²² The rule ($A \rightarrow B$) has the confidence measure of $P(B|A)$, which is the conditional probability of B given A. That is, $P(B|A)$ is computed by $P(A \cap B)/P(A)$. The last measure of ($A \rightarrow B$) is the lift. This measure is defined as the confidence measure divided by the probability of A. If the lift value of ($A \rightarrow B$) is 1, then A and B are independent of each other. The positive association between A and B can be strengthened by having the lift value exceed 1. In addition, a decrease in the lift value of A and B from 1 to 0 strengthens the negative association between A and B. For example, if the lift value of technology X and technology Y is 2, then there is a strong need to develop X and Y simultaneously. In this paper, data from 50,866 biotechnology-related patent documents retrieved from the USPTO were employed. The data was transformed into transaction data including IPC codes for ARM (Table 3).

For example, the second patent (patent 2) had developed technologies in technology fields C12N and A01H. From the transaction data, 688 ARM rules were extracted with minimum support = 0.001 and minimum confidence = 0.1. First, the top five meaningful rules were obtained for support, confidence, and lift measures (Table 4).

This result is somewhat different compared to the previous result⁸ obtained by the authors because the rules were extracted on the minimum support and confidence values in this paper. In other words, the rules were extracted by each measure of support, confidence, and lift, respectively, in the previous study. But, this research got the rules by considering both support and confidence measures to get more objective results.

It was found that the rules ($C12P \rightarrow C12N$) and ($C12P \rightarrow C12Q$) were meaningful because their

support and confidence values were ranked among the top five and determined the rules ($B01L \rightarrow C12M$) and (any IPC code $\rightarrow A61K$) as meaningful based on the lift measure. To determine more detailed relationships between IPC codes, a cluster analysis was conducted using IPC code data. A sample of 2,000 transactions was considered from 50,886 transactions for IPC code data clustering. In addition, the silhouette measure and partitioning around medoids (PAM) were employed.²³⁻²⁴ The silhouette measure was used to determine the number of clusters.²³ An optimal number of clusters was determined when the silhouette value of a selected number of clusters exceeded that of others.²³ In addition, the PAM was used to cluster IPC codes. For this, the R language and packages were used.²⁴⁻²⁶ The sampled IPC codes were divided into five clusters (average silhouette width = 0.45). The silhouette value of 0.45 was determined to be an appropriate criterion for selecting the number of IPC clusters and thus the sampled IPC codes were clustered into five groups by PAM (cluster size = 5) (Table 5).

Clusters 1, 2, 4, and 5 each included many transactions (patents), indicating that these clusters generally included developed technologies. Cluster 3 had fewer transactions, but the silhouette value for cluster 3 was higher than that for any other cluster. That is, cluster 3 was the most meaningful group. Thus, it was determined that the technologies in

Table 4 – Top five meaningful rules

Measures	Extracted rules
Support	($C12P \leftrightarrow C12N$), ($C07H \leftrightarrow C12N$), ($C07H \leftrightarrow C12Q$), ($C07H \leftrightarrow C12P$), ($C12P \leftrightarrow C12Q$)
Confidence	($A01H \rightarrow C12N$), ($\{C07H, C12M, C12P\} \rightarrow C12Q$), ($C11D \rightarrow C12N$), ($\{C07H, C12M\} \rightarrow C12Q$), ($A61F \rightarrow C12N$),
Lift	($B01L \rightarrow C12M$), ($\{A01N, C12P\} \rightarrow A61K$), ($\{A01N, C12Q\} \rightarrow A61K$), ($A01N \rightarrow A61K$), ($\{A01N, C12N\} \rightarrow A61K$)

Table 3 – Transaction data with IPC codes

Transactions	Items (IPC codes)
Patent-1	C12N, A01H
Patent-2	C12N, A01H, C07H
⋮	⋮
Patent-20	C12M, H01L, F27D, B01L, G05B
⋮	⋮
Patent-50885	C12Q
Patent-50886	C12N

Table 5 – IPC code data clustering

Clusters	Number of IPC codes	Silhouette values
1	619	0.63
2	473	0.39
3	95	0.70
4	237	0.40
5	576	0.12

Table 6 – Top five rules with consequent code = C12N

Antecedent	Consequent	Support	Confidence	Lift
A01H	C12N	0.0346	0.9832	1.6609
A01N	C12N	0.0127	0.7642	1.2909
C07K	C12N	0.0560	0.7405	1.2508
A61K	C12N	0.0448	0.7152	1.2081
C12R	C12N	0.0145	0.7023	1.1863

Table 7 – Top five rules with consequent code = C12Q

Antecedent	Consequent	Support	Confidence	Lift
C07H,C12M	C12Q	0.0065	0.9242	2.5850
C07H,G01N	C12Q	0.0129	0.8898	2.4889
G01N	C12Q	0.0662	0.8296	2.3205
C12P,G01N	C12Q	0.0101	0.7529	2.1061
C12M,G01N	C12Q	0.0052	0.6230	1.7425

cluster 3 were undeveloped because it was smaller than the others. In this paper, this cluster is referred to as a vacant technology area. Technologies belonging to vacant technology areas need to be developed in the future. If a firm develops a technology in a vacant technology area, then it gains a competitive advantage over other firms. The top five IPC codes were determined by their relative frequency in cluster 3: C07H, C12N, C12P, C12Q, and G01N. As shown in Tables 4 and 5, the antecedent IPC codes C12N and C12Q were meaningful for biotechnology forecasting. Thus, interesting rules were extracted whose consequent IPC codes included C12N and C12Q (Tables 6 and 7).

According to association rule theory, if technologies belonging to A01H are developed, then those belonging to C12N are also developed. Thus, the relationship between A01H and C12N was strong and positive based on their lift value. In addition, the technologies belonging to these IPC codes accounted for approximately 3.5 per cent of all biotechnology patents. Like A01H, A01N, C07K, A61K, and C12R had strong relationships with C12N.

Among the extracted rules with C12Q, the ARM rule ($\{C07H,C12M\} \rightarrow C12Q$) had the highest lift value. That is, if the technologies belonging to C07H and C12M are developed, then those belonging to C12Q must be developed. The rest of the rules can be explained in a similar manner.

Patent Clustering for Biotechnology Forecasting

For patent document clustering, the K-means clustering algorithm²² was employed, which is a popular method for document clustering.²⁷ K-means clustering entails the determination of clusters and their centres.²¹ In addition, the number of cluster

Table 8 – K-means clustering result (K=5)

Clusters	Number of patents	Percentage (%)
1	134	13.4
2	121	12.1
3	389	38.9
4	41	4.1
5	315	31.5
Total	1,000	100.0

Table 9 – High-frequency terms in cluster 4

Frequency rank	Extracted terms
1 st	Cell, cysteine, DNA, nucleic acid, polypeptide, protein
2 nd	Gene
3 rd	β -glucuronidase, mixture
4 th	Bacterial
5 th	Saliva, disease, molecules, cloning

centres (K) is determined first. This clustering procedure changes cluster centres to minimize the total within-variance of clusters iteratively.²¹ In this study, abstracts of 1,000 patent documents obtained from 50,886 patent documents were used. By preprocessing text data, a matrix of documents and terms in which rows and columns represented documents and terms, respectively, was constructed. That is, the element of the matrix was the observed frequency of a term in a patent document. The dimension of the constructed document-term matrix was (1,000)*(10,699), and thus, the matrix was too large to be analyzed through the K-means clustering algorithm. To reduce the dimension, a principal component analysis (PCA) was conducted.²⁸ PCA is the most widely used method for reducing dimensions.²² Through the PCA, the document-term matrix was transformed into a document-PC score matrix whose dimension was (1,000)*(426). The cumulative proportion of the document-PC score matrix was 90.01 per cent. The dimension of the column was reduced from 10,699 to 426 with about a 10 per cent loss of information. This matrix was then used for K-means clustering. As discussed earlier, the number of clusters based on the silhouette measure was five, and thus, K determined as 5. The clustering provided four clusters with large numbers of patents and one cluster with a relatively small number of patents (Table 8).

Cluster 4 was identified as a vacant technology area for biotechnology forecasting. This cluster was smaller than the others, that is, the technologies belonging to cluster 4 were relatively undeveloped. To group the technologies belonging to cluster 4, high-frequency

terms were extracted from this cluster (Table 9).

The results indicate that the vacant technology area included technologies reflecting terms such as 'cell', 'cysteine', 'DNA', 'nucleic acid', and 'polypeptide', among others.

Combining Results from the Three Approaches

In this paper, the following three approaches to biotechnology forecasting were examined: time series analyses, ARM, and patent clustering (Fig. 3).

The retrieved patent documents from USPTO were analysed by three approaches for TF in this paper and the three results for biotechnology forecasting combined. Next, the yields of the

biotechnology forecasting are used for patent management. One can predict the future demand of biotechnology using time series analysis and assign R&D resources of a company for developing biotechnology. IPC code analysis helps in identifying relationships between them. This provides the key patents in biotechnology domain. Also, one can develop or buy the technologies of key patents for biotechnology. Lastly, one can discover the vacant areas of biotechnology from the result of patent clustering and also detect the emerging and disruptive technologies in the field of biotechnology. Therefore, this research consists of two stages, technology forecasting and patent management.

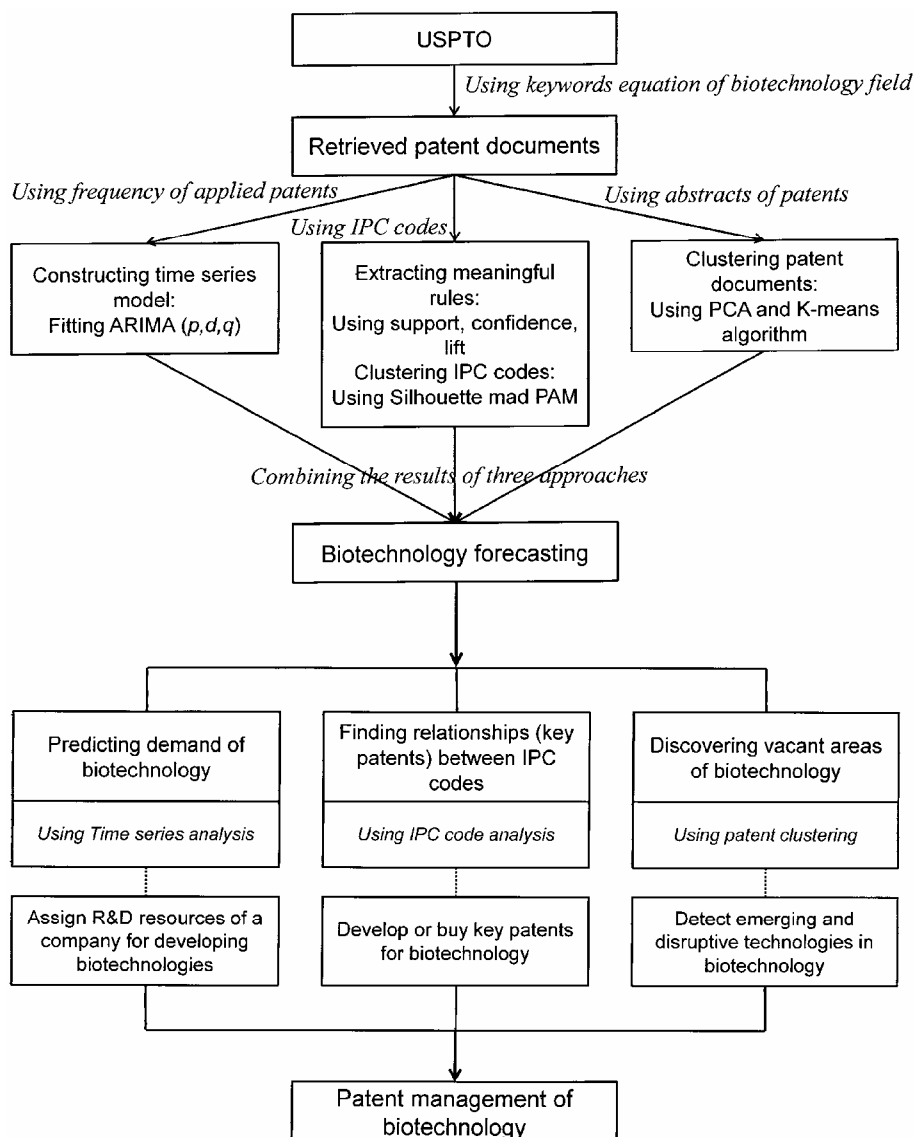


Fig. 3 – Biotechnology forecasting and patent management

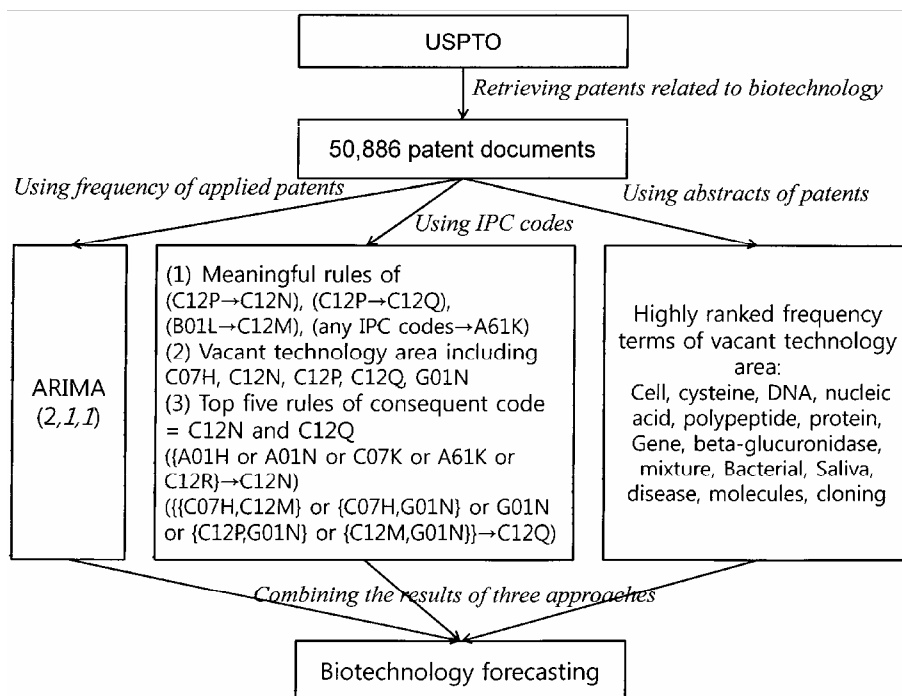


Fig. 4 – Combined results for three approaches to biotechnology forecasting

For the biotechnology-patent management, the valuable and meaningful results of the biotechnology forecasting have to be used. Using this process, forecasting results for patent management of biotechnology were obtained (Fig. 4).

As discussed earlier, the ARIMA (2,1,1) model was considered as the fitted model for forecasting the future frequency of biotechnology-related patent applications and determined that the technologies belonging to C12P were important antecedent technologies. The IPC code C12P represents technologies entailing ‘fermentation or enzyme-using processes to synthesis a desired chemical compound or composition or to separate optical isomers from a racemic mixture.’²³ The technologies belonging to C12N and C12Q were important consequent technologies. The IPC codes C12N and C12Q indicate technologies entailing ‘micro-organisms or enzymes; compositions; propagating, preserving, or maintaining micro-organisms; mutation or genetic engineering; culture media’ and ‘measuring or testing processes involving enzymes or micro-organisms; composition or test papers; processes of preparing such compositions; condition-responsive control in microbiological or enzymological processes,’ respectively.²³ Based on the results of patent clustering, the following high-frequency terms for vacant technology areas were obtained: ‘cell’,

‘cysteine’, ‘DNA’, ‘nucleic acid’, ‘polypeptide’, ‘protein’, ‘gene’, ‘beta-glucuronidase’, ‘mixture’, ‘bacterial’, ‘saliva’, ‘disease’, ‘molecules’, and ‘cloning’. Based on these terms, the necessary technologies were determined.

Conclusion

This paper proposes a patent management method for biotechnology forecasting. For this, three approaches were examined: the ARIMA model, ARM, and K-means clustering. A time series analysis using the ARIMA model was conducted and a model for predicting the future frequency of issued patents determined. In addition, meaningful association rules were extracted along with relationships between biotechnology-related IPC codes based on the results of ARM using IPC data. Finally, clustering results for patent documents were obtained using the silhouette measure and the K-means algorithm. By combining the results for the three approaches, vacant (necessary but undeveloped) technologies were forecasted. This method of study can be applied not only to biotechnology but also to other technology fields, and thus, the study contributes to research on patent management for TF.

A business person can develop a R&D strategy using the results of this research. For example, from the result of patent clustering, the high-frequency

terms for vacant technology areas of biotechnology were obtained. But, defining detailed technology based on these terms is the role of domain experts. By identifying vacant technology, one can carry out patent administration to avoid patent suits and save enormous royalties from this result. In other words, it is good to buy or develop the patents defined as vacant technology in advance.

However, this study has a limitation: domain knowledge of biotechnology was required to apply the results to biotechnology forecasting. Thus, future research should develop a model that is less dependent on domain knowledge for TF by using diverse data-mining and management-engineering techniques.

References

- Chang J and Zhu X, Bioinformatics databases: Intellectual property protection strategy, *Journal of Intellectual Property Rights*, 15 (6) (2010) 447-454.
- Thumm N, Management of intellectual property rights in European biotechnology firms, *Technological Forecasting & Social Change*, 67 (2-3) (2001) 259-272.
- Wydra S, Production and employment impacts of biotechnology-input-output analysis for Germany, *Technological Forecasting & Social Change*, 78 (2011) 1200-1209.
- Jun S, A forecasting model for technological trend using unsupervised learning, *Communications in Computer and Information Science*, 258 (2011) 51-60.
- United States Patent and Trademark Office (USPTO), 2011, <http://www.uspto.gov> (1 October 2011).
- Jun S, IPC code analysis of patent documents using association rules and maps - Patent analysis of database technology, *Communications in Computer and Information Science*, 258 (2011) 21-30.
- Jun S and Uhm D, Technology forecasting using frequency time series model: Bio-technology patent analysis, *Journal of Modern Mathematics and Statistics*, 4 (3) (2010) 101-104.
- Jun S, A new patent analysis Using Association Rule Mining and Box-Jenkins Modeling for technology forecasting, *Advanced Science Letters*, (2012) (in press).
- Patent Searching - Tools & Techniques*, edited by D Hunt, L Nguyen and M Rodgers M, (John Wiley & Sons, New Jersey), 2007, p. 46-47.
- World Intellectual Property Organization (WIPO), 2011, www.wipo.org (1 October 2011).
- Gopalan R, Bioinformatics: Scope of intellectual property protection, *Journal of Intellectual Property Rights*, 14 (1) (2009) 46-51.
- Roper A T, et al., *Forecasting and Management of Technology* (John Wiley & Sons, New Jersey), 2011, p. 40-46.
- Mahajan M M, The emergence of new R&D paradigms in the Indian pharmaceutical industry: Post TRIPS period, *Journal of Intellectual Property Rights*, 16 (4) (2011) 321-329.
- Jun S et al., Forecasting vacant technology of patent analysis system using self organizing map and matrix analysis, *Journal of the Korea Contents Association*, 10 (2) (2010) 462-480.
- Jun S and Uhm D, Patent and statistics, What's the connection? *Communications of the Korea Statistical Society*, 17 (2) (2010) 205-222.
- Jeon J, Lee C and Park Y, How to use patent information to search potential technology partners in open innovation, *Journal of Intellectual Property Rights*, 16 (5) (2011) 385-393.
- Watkins J M and Taylor M Z, Intellectual property protection and US foreign direct investment in emerging economies, *Journal of Intellectual Property Rights*, 15 (6) (2010) 415-428.
- Box G E P, Jenkins G M and Reinsel G C, *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, New Jersey), 2008, p. 170-177.
- Bowerman B L, O'Connell R T and Koehler A B, *Forecasting, Time Series, and Regression - An Applied Approach*, 4th edn (Brooks/Cole, US), 2005, p. 338-339.
- http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml (15 January 2012).
- Hastie T, Tibshirani R and Friedman J, *The Elementary Statistical Learning Data Mining, Inference, and Prediction* (Springer, New York), 2001, p. 437-444.
- Han J and Kamber M, *Data Mining Concepts and Techniques* (Morgan Kaufmann, San Francisco), 2001, p. 226-230.
- Rousseeuw P J, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 20 (1) (1987) 53-65.
- R: A language and environment for statistical computing, R Development Core Team, R Foundation for Statistical Computing, 2011, <http://www.R-project.org> (15 September 2011).
- Hahsler M and Hornik K, Building on the arules infrastructure for analyzing transaction data with R, *Proceedings of the 30th Annual Conference of Advanced in Data Analysis* (Freie Universität Berlin, Berlin), 2007, p. 449-456.
- Hahsler M, Graun B and Hornik K, arules - A computational environment for mining association rules and frequent item sets, *Journal of Statistical Software*, 14 (15) (2005) 1-25.
- Giudici P, *Applied Data Mining, Statistical Methods for Business and Industry* (John Wiley & Sons, England), 2003, p. 121-126.
- Johnson R A and Wichern D W, *Applied Multivariate Statistical Analysis*, 3rd edn (Prentice Hall), 1992, p. 357-379.