New Technology Management Using Time Series Regression and Clustering

Sang Sung Park\(^1\) and Sunghae Jun\(^2\)*

\(^1\)Division of Information Management Engineering, Korea University, Seoul, Korea
\(^2\)Department of Statistics, Cheongju University, Cheongju, Korea
hanyul@korea.ac.kr, shjun@cju.ac.kr

*Corresponding Author: Sunghae Jun* (shjun@cju.ac.kr)

Abstract

Technology can be defined diversely according to the ways of its usage. In this paper, we define technology as a tool making new product and service using the developed results of science and engineering for improving the quality of human life. Technology management (TM) is an important factor in the business planning of a company. Many companies have performed TM for developing new products and protecting their intellectual properties (IP). Patent is a typical IP, so we propose a TM approach using new patent analysis method. In this paper, we combine time series regression and clustering techniques. To assess the performance of our research, we will make experiment using the biotechnology patent data from the United State Patent and Trademark Office.

Keywords: Technology management, patent analysis, time series model, regression analysis, clustering

1. Introduction

Market, product, and technology are important issues in a company [1]. Among them, technology is more important than others because technology can push and pull the market and product [1]. This research defines technology as a tool making new product and service using the developed results of science and engineering for improving the quality of human life. Also, technology management (TM) is an important factor in the business planning of a company [2]. Many companies have performed TM for developing new products and protecting their intellectual properties (IP). Patent is a typical IP, so we propose a TM approach using new patent analysis method. This approach is based on technology forecasting (TF). TF is to predict the future state of a technology field [3-5] and an important factor in TM. The R&D policy is depended on the result of TF. Patent analysis (PA) has performed an important role in the TF [6]. So, this research proposes a PA method for TM. We will combine the time series regression and clustering methods for constructing PA model in this paper. PA is a popular area in the TM and TF fields [7]. We also researched and developed some PA approaches for TF [8-9]. Most of the published PA methods were depended on an analytical approach such as supervised and unsupervised learning methods or considered combining the results of some PA methods. But, this research combines the PA methods directly. In other words, we get together not the results of a couple of PA methods but the PA methods themselves. In this paper, we consider time series regression and document clustering as supervised and unsupervised learning approaches respectively. So, this research constructs a TF model for an efficient TM. To assess our improved performance, we will
make experiment using the biotechnology patent data from the United State Patent and Trademark Office (USPTO) [10].

2. Patent Analysis and Technological Forecasting

Patent is an IP containing the entire information of developed technologies [4]. This has the exclusive right for the inventor’s technology for a limited period. A patent document contains the important source of information for the technology such as title, abstract, description of the invention, claims, drawings, issued date, citations and so on [3]. So, to know the trend of a technology, we have to analyze the patent documents. PA is to analyze the patent documents data. PA plays a major role in R&D policy because we can forecast the future aspect of a technology by the result of PA. So, most companies make efforts to perform PA for improving their competitiveness. Traditional PA methods were based on the quantitative analyses such as statistics and machine learning. In this paper, we consider clustering and regression approaches together for efficient PA. Also, we construct a TF model from the result of this PA. Finally, we build the proposed TM structure using the TF model.

3. New Technology Management Model

This research proposes a TF model for efficient TM. This has two analytical methods, supervised and unsupervised learning approaches which are time series regression and K-means clustering algorithm. In this paper, we combine two methods for efficient TF. In the document data clustering [11-12], K-means clustering is a popular non-hierarchical clustering method for finding K groups and assigning all data to each cluster [13]. In this paper, we use this clustering method because patent data are also document data. With the K-means clustering method as unsupervised learning, we use time series regression as supervised learning. Time series regression is a forecasting model to construct the function of response variable Y and explanatory variable X as follows [14].

\[ y_i = b_0 + b_1 x_i + \varepsilon_i \]

Where \( y_i \) is the number of issued patent documents in time period \( i \). \( x_i \) and \( \varepsilon_i \) are the time period \( i \) and the error term respectively. Also, \( b_0 \) and \( b_1 \) are regression parameters. For the assessment of this model, we use the coefficient of determination (R-square) and probability value (p-value) of the regression parameter. Figure 1 shows the entire process of our developed approach.

![Figure 1. New PA Model for TF](image-url)
First, we select the technology field for TF and retrieve the patent documents related to the selected technology. At this moment, we use the keywords equation of the technology. The retrieved patent documents are transformed to structured data for K-means clustering because this method needs the structured data as input data. In this paper, we preprocess the patent document data by text mining technique. Text mining is to analyze the text documents [15]. From the titles and abstracts of the patent documents, we construct a document-term matrix as the structured data using the text mining based on text corpus, parsing, and text database. The document-term matrix consists of documents and terms. A row of this matrix represents a document and a column shows a term. Each value of the matrix is the occurred frequency of a term in a document. We perform K-means clustering using this matrix data (structured data). The result of K-means clustering provides K clusters. Also, we construct K time series regression models from the result of K-means clustering. Figure 2 shows this process in detail.

![Figure 2. TF Process in Cluster j](image)

Cluster \( j \) is the \( j \)-th cluster of \( K \) clusters. We define the representative technology of cluster \( j \) using the extracted keywords from the patent documents assigned in cluster \( j \). We also predict the technological trend of cluster \( j \). For this, this research construct time series regression model using the numbers of issued patents in cluster \( j \) by year. By this process, we decide whether the representative technology of cluster \( j \) is emergence or not. In this paper, we determine the technology with highly increasing trend to emergence-technology. The statistical decision of increasing degree is depended on the probability value (p-value) of the parameter of time series regression. Using p-value measure, we get the emergence-technology from the \( K \) technology areas of \( K \) clusters.

4. Experimental Results

To verify our improved performance, we retrieved the biotechnology patent documents from USPTO [10]. We searched total 50,886 patent documents until 2009. To construct the proposed model, we used 1,500 random samples from the searched patent document data. First, we performed K-means clustering using this data set. The \( K \) was selected to 4 because biotechnology consisted of four international patent classification (IPC) codes which were C12M, C12N, C12P, and C12Q [16]. We used R language, function and package for this experiment [17-18]. Table 1 shows the result of K-means clustering.

**Table 1. K-means Clustering Result of Biotechnology Patent Documents**

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Number of assigned patents</th>
<th>Within cluster sum of square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>529</td>
<td>1253.85</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>1369.58</td>
</tr>
<tr>
<td>3</td>
<td>211</td>
<td>1242.65</td>
</tr>
<tr>
<td>4</td>
<td>723</td>
<td>10146.88</td>
</tr>
</tbody>
</table>
We knew the detailed biotechnologies represented by cluster 1 and 3 had been developed widely because the numbers of patent documents of cluster 1 and 3 were larger than other clusters. But, cluster 2 had a relatively small number of patents. So, we knew the technology represented by cluster 2 had been undeveloped until now. Next, to define the representative technologies of four clusters, we extracted top ten keywords from the clusters in Table 2.

### Table 2. Top Ten Keywords of Four Clusters

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Top ten keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>amplification, rna, dna, cell, recombinant, catalytic, nucleic, synthase, protein, amino</td>
</tr>
<tr>
<td>2</td>
<td>particles, flow, concentration, target, material, member, support, binding, channel, sequence</td>
</tr>
<tr>
<td>3</td>
<td>receptor, chamber, reaction, serum, hematopoietic, neural, functional, splice, urological, myocardial</td>
</tr>
<tr>
<td>4</td>
<td>gramineae, human, mammalian, taurine, plant, infected, seedling, hydrogen, stem, cultivation</td>
</tr>
</tbody>
</table>

In the process of extracting top ten keywords from the patent documents assigned in four clusters, we removed the meaningless terms such as ‘and’, ‘for’, ‘the’, and so on. The detailed technology of cluster 4 could be defined by the ten terms which were ‘gramineae’, ‘human’, ‘mammalian’, ‘taurine’, ‘plant’, ‘infected’, ‘seedling’, ‘hydrogen’, ‘stem’, and ‘cultivation’. So, in this paper, we defined the representative technology of cluster 4 as the biotechnology related to agriculture and farming. Table 3 shows the defined technologies of four clusters.

### Table 3. Representative Technologies of Four Clusters

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Representative technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Basic features of living organism</td>
</tr>
<tr>
<td>2</td>
<td>Interaction between living things</td>
</tr>
<tr>
<td>3</td>
<td>Functions of immunity and sense for living</td>
</tr>
<tr>
<td>4</td>
<td>Biotechnology related to agriculture and farming</td>
</tr>
</tbody>
</table>

Cluster 1 contains the technology of basic features of living organism in biotechnology. The representative technology of cluster 2 is the technology of interaction between living things. The technology of Functions of immunity and sense for living is included in cluster 3. Cluster 4 consists of the technology of biotechnology related to agriculture and farming. Figure 3 shows the trends of the representative technologies of four clusters by year.

![Figure 3. Technological Trends of Four Clusters](image-url)
The number of issued patent documents including each cluster is shown in Figure 3. The technological trend of cluster 2 was different to other clusters. We knew the cluster 3 and 4 had similar technological trends. Using the time series data of four clusters, we constructed the TSR models. Table 4 represents the result of four TSR models.

### Table 4. Time Series Regression Results of Four Clusters

<table>
<thead>
<tr>
<th>Clusters</th>
<th>R-square (%)</th>
<th>Beta</th>
<th>p-value</th>
<th>F-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.21</td>
<td>1.04</td>
<td>0.0006</td>
<td>15.16</td>
<td>0.0006</td>
</tr>
<tr>
<td>2</td>
<td>2.10</td>
<td>56.83</td>
<td>0.5007</td>
<td>0.44</td>
<td>0.5128</td>
</tr>
<tr>
<td>3</td>
<td>33.42</td>
<td>-857.84</td>
<td>0.0014</td>
<td>13.06</td>
<td>0.0013</td>
</tr>
<tr>
<td>4</td>
<td>24.36</td>
<td>-19.31</td>
<td>0.0045</td>
<td>9.66</td>
<td>0.0041</td>
</tr>
</tbody>
</table>

The model of cluster 2 was poor because the coefficient of determination (R-square) was very small and the probability (p-value) is larger than 0.05. Other clusters had significant models. Therefore, in this paper, we determined the technology of cluster 3 was emerging area for biotechnology because the number of issued patent documents of cluster 3 was relatively small and its model was statistically significant. From this TF result, we conclude that the technologies related to ‘functions of immunity and sense for living’ will be needed for biotechnology field. Also, the future trend of this technology has the following model.

\[ Y = -857.84 + 0.0014X \]

How to plan efficient and new TM strategy based on our result is a role of the biotechnology experts.

### 5. Conclusions

In this paper, we proposed a TM model using patent clustering and time series regression. We used K-means clustering method for patent document clustering. Also, we constructed the technological trend model of each cluster using time series regression. The aim of this paper was to combine two PA methods which are K-means clustering and time series regression effectively for TF and TM. In our work, the retrieved patent documents were transformed into structured data using text mining techniques for the quantitative analysis. Using the structured data, we performed our combined model to find emerging area of biotechnology. We determined this area from the results of patent clustering and time series regression. In the clustering result, we decided the technology of the cluster with relatively small size to the emerging technology. Also, we verified its significance by the result of time series regression model. This research contributes to a TF domain to search the emerging technology for efficient TM. But, our work had a limitation which was to select the emerging technology subjectively. In other words, we need the domain expert’s knowledge to define the technology of the clusters. In our future work, we will develop more objective TF approach to find the emerging technology.

### References


Authors

**Sang Sung Park**

He received the MS and PhD degrees in department of Industrial System and Information Engineering at Korea University, and his the BS degree in department of Industrial Engineering at Dongyang University. He is a Research Professor at Korea University. His research interests include intellectual property, data mining and pattern classification.

![Sang Sung Park](image)

**Sunghae Jun**

He received the BS, MS, and PhD degrees in department of Statistics, Inha University, Korea, in 1993, 1996, and 2001. Also, He received PhD degree in department of Computer Science, Sogang University in 2007. He was a visiting scholar in department of Statistics, Oklahoma State University, USA in 2009-2010. He is currently Associate Professor in department of Statistics, Cheongju University, Korea. He has researched data mining and management of technology (MOT).

![Sunghae Jun](image)