A Statistical Text Mining Method for Patent Analysis

Sunghae Jun

Department of Statistics Cheongju University, shjun@cju.ac.kr

Abstract

Most text data from diverse document databases are unsuitable for analytical methods based on statistics and machine learning algorithms. Patent documents are also compiled into text datasets. Similar to other document datasets, we therefore need to transform patent documents into structured data for a statistical analysis. This transformation is performed using the preprocessing of text mining techniques. We can analyze the patent documents after their preprocessing. For a patent analysis, two phases, preprocessing and analysis, are required. In this paper, we try to combine the two phases into one. We propose a statistical text mining method to improve the performance of a patent data analysis. Our proposed method carries out text mining and a statistical analysis at the same time. To show the contribution of our study, we illustrate how it can be applied in a real domain using a target technology.

Keywords: Statistical text mining, Preprocessing, Statistical analysis, Patent system, Technology forecasting, Patent analysis

1. Introduction

A patent is a type of intellectual property (IP). Unlike IPs such as trademarks, copyrights, and trade secrets, a patent includes diverse information of technological development [1]. In addition, the patent system protects the exclusive rights of inventors regarding their developed technology for limited periods of time [2]. Many companies hope to analyze patent documents to ascertain certain technology trends. However, because patent data include text, numbers, dates, and pictures, patent data are unsuitable for analysis methods such as statistics and machine learning algorithms [1], which require structured data consisting of numbers or frequencies. To solve this problem, we should preprocess patent documents using text mining techniques [3]. Text mining is a data mining process used to manage and analyze texts or document data [4]. We therefore use text mining methods to preprocess patent data to build structured data for a statistical analysis. In this paper, we combine statistics and text mining for a patent analysis. This type of combination is called statistical text mining (STM) and has been used in medical and bio sectors [5][6]. This research provides another STM approach for patent analysis, and differs from previous studies. In our STM, we used descriptive statistics and a multivariate analysis based on statistics. In addition, using text mining techniques, we import data, create a corpus, and eliminate whitespace and stop words. In other words, we apply basic and advanced analyses to structured patent data after the preprocessing of patent documents. The STM results can be used for R&D planning, technology management, or new product development. To show how our STM can be used in a real domain, we perform a case study using retrieved patent documents from the Korea Intellectual Property Rights Information Service (KIPRIS) database. We then apply our STM process to this case study in a step-by-step manner.

2. Statistical and text mining

Statistics and text mining are both data mining tools [4]. Statistics is a leaning method for prediction and forecasting [7]. Most statistical methods are based on a probability distribution such as a binomial or normal distribution. Text mining is an analytic process for discovering novel information from a large amount of text data [4]. Differing from general data mining, text mining requires some natural language process techniques because the preprocessing of text data should be performed to analyze text data.
3. Statistical text mining for a patent analysis

Statistics plays an important role in big data analyses. Patent documents are typical examples of big data. A patent document contains diverse data types such as text, numbers, dates, and pictures [1]. Patents include a lot of information regarding the results of technological developments. Most companies have therefore tried to analyze patent data for their technology management such as R&D planning and the development of new products. Many statistical methods have been used for patent analyses [8][9]. However, certain problems exist in patent analyses using statistical methods. One such problem is that patent documents are not suitable for statistical analyses because traditional statistics requires numerical data. Several studies to overcome this problem have therefore been conducted [8][10]. Text mining is a good approach to settle these obstacles [3]. Therefore, we combine a statistical analysis and text mining, called statistical text mining, for a patent analysis. Figure 1 shows the scope of the proposed statistical text mining method.

![Interdisciplinary statistical text mining](image)

**Figure 1. Interdisciplinary statistical text mining**

We researched STM for use in a big data analysis, and analyzed patent documents as examples of big data. Statistics and text mining are different fields used in big data analyses. Statistics is a natural science based on probability for estimation, prediction, or forecasting. Text mining is an applied information science based on computer science for text data analyses. Table 1 shows a comparison between statistics, text mining, and statistical text mining based on their respective scopes [4][7].

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Statistical text mining</th>
<th>Text mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive statistics</td>
<td>Structured data construction</td>
<td>Information retrieval</td>
</tr>
<tr>
<td>- sample statistic</td>
<td>- importing text data</td>
<td>- precision &amp; recall</td>
</tr>
<tr>
<td>- tables &amp; graphs</td>
<td>- creating corpus</td>
<td>- document selection</td>
</tr>
<tr>
<td>Probability</td>
<td>- building document term matrix</td>
<td>Text indexing</td>
</tr>
<tr>
<td>- support and confidence</td>
<td>Preprocessing</td>
<td>- latent semantic indexing</td>
</tr>
<tr>
<td>- Bayes’ rule</td>
<td>- eliminating whitespace</td>
<td>- locality processing indexing</td>
</tr>
<tr>
<td>Random variable</td>
<td>- removing stop and common words</td>
<td>- probabilistic indexing</td>
</tr>
<tr>
<td>- discrete &amp; continuous Probability distribution</td>
<td>Basic analysis</td>
<td>Query processing</td>
</tr>
<tr>
<td>- binomial &amp; Poisson</td>
<td>- summary statistics</td>
<td>Document classification</td>
</tr>
<tr>
<td>- normal</td>
<td>- meta analysis</td>
<td>- document ranking</td>
</tr>
<tr>
<td>Inference</td>
<td>Advanced analysis</td>
<td>Web mining</td>
</tr>
<tr>
<td>- estimation</td>
<td>- finding association rules</td>
<td>- web usage mining</td>
</tr>
<tr>
<td>- hypothesis testing</td>
<td>- regression &amp; path analysis</td>
<td>- web structure mining</td>
</tr>
<tr>
<td></td>
<td>- text or document clustering</td>
<td>- web recommendation system</td>
</tr>
</tbody>
</table>

Our STM is focused on patent documents. The statistics field in Table 1 is some distance from text mining. Statistics is used to summarize and visualize numeric data, and to infer a population based on estimation and hypothesis testing. In comparison, text mining is used to process text data through retrieval and indexing. However, STM makes structured data suitable for a statistical analysis. The structured data construction and preprocessing of STM are based on text mining techniques. In addition,
the basic and advanced analyses of STM are based on statistical methods. Figure 2 shows the step-by-step process of our proposed STM method.

4. Experimental results

For our case study, we searched for patent documents from the KIPRIS patent database [11]. The keyword equation is as follows: title = text * mining; in other words, we retrieved patent documents related to text mining technology. In addition, we obtained patents from the United States and Europe on July 5, 2013. The patent data consisted of 67 documents in total. Among them, 60 were US patents, and the remaining were Europe patents. Next, we illustrate the case study for STM in a step-by-step manner.

4.1. Structured data construction

In this paper, we performed text mining techniques to construct structured data for a statistical analysis. To preprocess the patent data used in our research, we used the text mining package of the R project [12][13]. We downloaded the patent data as an Excel file and imported it into the R project. We created a corpus of 67 text documents. From this corpus, we constructed a patent term matrix using 67 documents and 1439 terms. We found the following highly frequent (over 10) terms in the structured data: “also,” “analysis,” “analyzed,” “and,” “apparatus,” “are,” “associated,” “based,” “between,” “business,” “can,” “candidate,” “characteristic,” “coincident,” “computer,” “condition,” “confidence,” “content,” “corresponding,” “data,” “definition,” “degree,” “device,” “disclosed,” “document,” “each,” “element,” “entities,” “extracted,” “extracting,” “extraction,” “extract,” “feature,” “for,” “frequency,” “from,” “generating,” “generation,” “has,” “include,” “including,” “information,” “inherent,” “input,” “into,” “knowledge,” “language,” “least,” “list,” “may,” “means,” “method,” “mining,” “more,” “obtained,” “one,” “phrase,” “piece,” “plurality,” “portion,” “predetermined,” “present,” “processing,” “program,” “provided,” “query,” “referring,” “result,” “said,” “search,” “section,” “set,” “structured,” “such,” “system,” “term,” “text,” “textual,” “that,” “the,” “this,” “topic,” “unit,” “used,” “user,” “using,” “value,” “video,” “web,” “which,” “with,” “within,” “word,” and “words.”
To build the patent keyword matrix, we selected keywords from the frequent terms by eliminating stop and common words. In this way, we determined the following candidate keywords: “analysis,” “associated,” “business,” “candidate,” “characteristic,” “coincident,” “computer,” “condition,” “confidence,” “content,” “data,” “definition,” “degree,” “device,” “disclosed,” “document,” “element,” “entities,” “extract,” “feature,” “frequency,” “generation,” “including,” “information,” “inherent,” “input,” “knowledge,” “language,” “list,” “means,” “method,” “mining,” “phrase,” “piece,” “plurality,” “portion,” “predetermined,” “present,” “processing,” “program,” “query,” “referring,” “search,” “section,” “set,” “structured,” “system,” “term,” “text,” “topic,” “unit,” “used,” “user,” “value,” “video,” “web,” and “word.”

The final keywords can be determined based on the knowledge of experts in the text mining field. In this paper, we determined the top ten keywords from the candidates other than “text” and “mining.”

<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business, confidence, content, disclosed, extract, feature, information, query, search, structured</td>
</tr>
</tbody>
</table>

We therefore constructed a patent keyword matrix using 67 patents and 10 keywords. Using this matrix data, we performed basic and advanced analyses, the results of which were used for practical application.

4.2. Basic analysis

STM provides a summary and visualization of the searched patent data in the basic analysis step. The following figure shows the number of patents by year.

![Figure 3. Number of applied patents](image)

The first patent related to text mining was applied in 1999. We therefore can see that the technological development of text mining was started comparatively recently. The development of text mining technology has also recently decreased. In addition, we summarized the International Patent Classification (IPC) codes of the searched patent data. IPC is a patent classification system from the World Intellectual Property Organization [14]. The following table shows the IPC codes used in text mining patents and their representative technology.

<table>
<thead>
<tr>
<th>IPC</th>
<th>C12Q</th>
<th>G06F</th>
<th>G06K</th>
<th>G06N</th>
<th>G06Q</th>
<th>G06T</th>
<th>G10L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of IPCs</td>
<td>1</td>
<td>58</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

We found that the text mining patents consisted of seven IPC codes. Most technologies of these text mining patents are based on the IPC code, G06F, which describes the technology of “electric digital data processing” [14]. The IPC code, C12Q, represents the technology of “measuring or testing
processes involving enzymes or micro-organisms,” and the remaining IPC codes are related to information technologies. We therefore can see that text mining technology is an interdisciplinary area for developing technologies. Next, we attempted to find more detailed knowledge through an advanced analysis.

4.3. Advanced analysis

In this section, we applied a statistical analysis to the structured data. First, we performed the following correlation analysis:

We know that “confidence,” “feature,” and “structured” have relatively large correlation coefficients for “text.” In addition, “mining” has a relatively large correlation coefficient between “confidence,” and we therefore concluded that the development of text mining technology is needed for a confidence technology. In addition, we should consider featured and structured approaches for text mining technology development. We performed another STM analysis as follows.

We carried out multiple regression analysis, estimate and significance are regression parameter and probability value (p-value) respectively. We can find the relative influence of a keyword to text or mining using the estimated value, and statistical testing using the significance (less than 0.05). Similar to the results of a correlation analysis, we found that the confidence technology is important for the development of text mining technology. We next applied our results to a practical problem.
4.4. Practical application

Using ten extracted keywords, we found that the disclosure, extraction, query, and search of information and structured content are important for business problems. In addition, based on the results of an IPC code summary, we can see that text mining technology can be applied to bio-systems. The basic analysis results show that the technologies related to text mining are comparatively recent. Finally, based on the results of an advanced analysis, we obtained the statistical significance of confidence technology on text mining. Therefore, to build an R&D plan for text mining technology, we recommend the results of our step-by-step STM process.

5. Conclusion and future work

In this paper, we proposed an STM methodology for a patent analysis. The proposed model consists of a patent document search related to the target technology, structured data construction, basic and advanced analyses, and practical application. To illustrate how our STM can be applied to a real domain, we performed a case study using the patent data related to text mining technology. Our research can contribute to the effective R&D planning or technology management of a company. For future work, we will study the development of the STM process for a patent analysis in more detail.

Acknowledgement

This work was supported by a research grant from Cheongju University in 2012.

References