Expert Systems with Applications 42 (2015) 4348–4360

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Keyword selection and processing strategy for applying text mining to patent analysis

Heeyong Noh, Yeongran Jo, Sungjoo Lee *

Department of Industrial Engineering, Ajou University, San 5, Woncheon-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-749, South Korea

ARTICLE INFO

Article history:
Available online 2 February 2015

Keywords:
Patent analysis
Text-mining
Keyword selection
Keyword processing
Document clustering

ABSTRACT

Previous studies have applied various methodologies to analyze patent data for technology management, given the advances in data analysis techniques available. In particular, efforts have recently been made to use text-mining (i.e. extracting keywords from patent documents) for patent analysis purposes. The results of these studies may be affected by the keywords selected from the relevant documents – but, despite its importance, the existing literature has seldom explored strategies for selecting and processing keywords from patent documents.

The purpose of this research is to fill this research gap by focusing on keyword strategies for applying text-mining to patent data. Specifically, four factors are addressed; (1) which element of the patent documents to adopt for keyword selection, (2) what keyword selection methods to use, (3) how many keywords to select, and (4) how to transform the keyword selection results into an analyzable data format. An experiment based on an orthogonal array of the four factors was designed in order to identify the best strategy, in which the four factors were evaluated and compared through k-means clustering and entropy values. The research findings are expected to offer useful guidelines for how to select and process keywords for patent analysis, and so further increase the reliability and validity of research using text-mining for patent analysis.

1. Introduction

Patent documents include bibliographical information such as application date, filing date, assignees and inventors, as well as descriptions of the novelty of the invention and its application areas as covered by the corresponding patent (Yoon & Park, 2004). They have widely regarded as an important source for evaluating technological strength and weakness and/or corporate R&D efforts and performance (Li, Wang, & Hong, 2009), and the bibliographic information in patent documents have been widely used for technology analysis and management – e.g. identifying technology trends, predicting emerging technologies (Basberg, 1987), and assessing technological capabilities at individual, firm, sector and national levels (Ernst, 2003).

Technological information extracted from patent data – the descriptive element of patent documents – has also recently been utilized in various advanced data analysis techniques and in developing text-mining tools (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998; Murphy et al., 2014; Trippe, 2003): in particular, the automatic extraction of major keywords from patent documents has been applied in technology management contexts (Dou, Leveillé, Manullang, & Dou, 2005). Whereas some researchers are still skeptical about the effectiveness of patent analysis based on this keyword-based approach (Krier & Zacca, 2002), others have emphasized its value and potential. For example, Fattori, Pedrazzi, and Turra's study proved that patent classification using text-mining could be effective, and could also overcome the limitations of conventional patent classifications (Fattori, Pedrazzi, & Turra, 2003): other researchers have used text-mining to conduct patent analyses, and shown that the approach is valuable for creating new technology and identifying technology opportunities.

In keyword-based studies, researchers have commonly tried to achieve their study goals through analyses using sets of keywords extracted from patent documents (Yoon, Lee, & Lee, 2010). Analysis results will thus depend on the keyword set that is selected – if it does not represent the characteristics of the entire document well, the reliability and accuracy of the subsequent analysis may be affected, which in turn will make it difficult to draw reliable insights from the results. Thus selecting and processing keywords that represents the patent's key technological concepts accurately is critical but challenging in patent analysis as modeling bibliographic data is significant but challenging in bibliometric analysis

* Corresponding author. Tel.: +82 31 219 2419; fax: +82 31 219 1610.
E-mail addresses: nhy6692@ajou.ac.kr (H. Noh), mistylake2357@ajou.ac.kr (Y. Jo), sungjoo@ajou.ac.kr (S. Lee).

http://dx.doi.org/10.1016/j.eswa.2015.01.050
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

(Ferrara & Salini, 2012). The importance of keyword selection and processing has been recognized not only in the field of patent analysis research, but also in text-mining application (Cheong, Chiu, Shu, Stone, & McAdams, 2011; Clifton, Cooley, & Rennie, 2004; Li et al., 2009) – but despite its importance, few previous studies have dealt with the factors that affect effective keyword selection and processing for patent analysis. Most have assumed that the keywords used in their studies have been extracted well, and have not examined the keyword selection processes carefully: so a systematic investigation of keyword selection and processing strategy for patent analysis is badly needed.

This research, therefore, focuses on the keyword selection and processing strategy for applying text-mining to patent analysis, and proposes some relevant guidelines. The strategies commonly used in the existing literature are reviewed and a method developed to evaluate their performance. Based on this, the performance is evaluated and the best suggested. The research findings are expected to help in the effective strategic use of keywords for patent analysis and thus further increase the reliability and validity of future research applying text-mining to this end.

The overall structure of this paper is as follows. Section 2 describes the basic trends of text-mining based patent analysis, and Section 3 discusses four significant factors regarding keyword selection and processing strategies for patent analysis. Section 4 explains the overall research framework and the detailed research methods, and the research results are described in Section 5. Finally, Sections 6 and 7 present the implications and limitations of our research, together with some concluding remarks.

2. Text-mining based patent analysis

Text-mining and its applications have received a lot of attention as a method to acquire useful information from unstructured corpora. Text-mining applications can be utilized in various domains; i.e. not only to help novel thinking (Gentner & Markman, 1997; Segers & De Vries, 2003), but also to create artificial intelligence (Falkenhainer, Forbus, & Gentner, 1986; Salton & Waldstein, 1978). In addition, as more reliable tools are being developed for text analysis, it has become possible to capture useful text information for an analysis that was unavailable within conventional approaches (Fujii, Iwayama, & Kando, 2007; Mukherjea, Bamba, & Kankar, 2005; Trippe, 2003). Especially in recent days, text-mining approach is utilized actively in technology management fields. A specific application includes text-mining based patent analysis, where patents are analyzed to investigate technology characteristics. Patent documents are considered as a valuable database for understanding technology trends and establishing innovation strategy because of the four reasons. First, patent documents are fully opened to the public, being accumulated for each year and each technological field. They contain information about almost all relevant technological fields, and (although there are a few exceptions) the great majority of novel inventions are patented. Hence, if text-mining tools can extract technological contents effectively, patent databases can provide a valuable source for in-depth technology analysis. Thus we can use patent documents to investigate technological trends, assess technological capabilities, and analyze the commercial value of technologies (Choi & Hwang, 2014). Secondly, patent databases are easily accessible – the advancement of IT and patent database systems has made it easier to obtain patent documents by downloading them through the internet (Schwander, 2000). Thirdly, the descriptive parts of patent documents are written in natural language, but in the same formats with consistent headings. Patent data are semi-structured, rather than unstructured, and technological contents are relatively easy to extract using text-mining tools (Kang, Na, Kim, & Lee, 2007). Finally, patent database can be a way to resolve a chronic limitation of the text-mining approach. The limitation of keyword-based approaches is that keywords can have various meanings, so keyword-based analysis results may misrepresent facts. However, most of terms used in patent documents are technical in nature, making it more likely that keywords have only single meanings, so the problems associated with text-mining approaches are expected to be relatively less severe in patent data than other applications (Lee, Yoon, & Park, 2009). Cheong et al. (2011) argued that engineering (or technological) keywords are not always useful for representing documents' contents but this is not true for patent documents (Kang et al., 2007). That is, engineering (or technological) terms can be used as representative keywords of patent documents because of the unique characteristics of patents, as was mentioned above. In addition, a number of studies showed that a text-mining based patent analysis with WordNet or latent semantic analysis enables to construct word ontologies systematically by identifying synonyms or hypernyms–hyponyms of a set of keywords (Fu, Cagan, Kotovsky, & Wood, 2013; Fu, Chan, Cagan, Kotovsky, Schunn, & Wood, 2013; Mukherjea et al., 2005; Murphy et al., 2014; Verhaegen, D'hondt, Vandevenne, Dewulf, & Duflou, 2011). These four characteristics of the patent database make text-mining – whose main merits are comprehensiveness, standardization and general applicability – has become more widely utilized for patent analyses.

Previous studies applying text-mining to patents can be divided into three categories. First, there are patent-map related studies, which have suggested methods to map the technological characteristics of patent documents visually, so as to identify new technology opportunities (Kim, Suh, & Park, 2008; Kim et al., 2014; Li et al., 2009; Son, Suh, Jeon, & Park, 2012; Tseng, 2005) or even management opportunities such as M&A (Park, Yoon, & Kim, 2013). Other studies have addressed the relationships between patents by conducting network analyses based on keywords (Yoon & Park, 2004). Text-mining has been combined with other analysis methods – such as conjoint analysis and data envelopment analysis – to obtain more meaningful findings for analyzing technology trends and identifying new technologies (Daim, Rueda, Martin, & Gerdsri, 2006; Lee, Lee, & Yoon, 2011; Seol, Lee, & Kim, 2011; Yoon & Park, 2007). A second research stream has emphasized text-mining methods' ability to reduce the huge amounts of resources and efforts necessary for the technology classification of patent documents, not simply advancing patent classification techniques, but also proposing automatic classification systems for patent documents (Chakrabarti, Dom, Agrawal, & Raghavan, 1998; Fall, Törcsvári, Benzineb, & Karetka, 2003; Lamirel, Al Shehabi, Hoffmann, & François, 2003; Lee et al., 2009; Liang, Tan, & Ma, 2008; Trappey, Trappey, Hsu, & Hsiao, 2009; Tseng, Lin, & Lin, 2007), whose effectiveness has already been verified by many research institutes. Finally, there are a set of previous studies concerning how to extract meaningful keywords when a text-mining approach is applied to patent documents. These studies are again grouped into two types. Most of them have focused on meaningful keywords extraction as tools to solve a certain problem. For example, researchers tried to solve a TRIZ problem by constructing meaningful keyword ontology (Liang et al., 2008; Souili & Cavallucci, 2013; Souili, Cavallucci, Rousselot, & Zanni, 2011). On the other hand, a few others, though not many, have concentrated on the type of text-mining approaches that are appropriate for patent analysis. For example, researchers tried to compare several keyword selection criteria including keyword frequencies in documents, variances of keyword frequencies across documents, and TF–IDF values (Lee et al., 2009; Li et al., 2009), while others have sought to identify the most appropriate parts of patent documents from which to extract keywords, such as titles, abstracts, claims and descriptions (Xie & Miyazaki, 2013).

In spite of the value of these existing studies, most of them focus only on analysis methods and the results of using keywords, on the assumption that the keywords themselves were selected well enough to represent the contents of patent documents. The second type of the third research stream has been conducted to address this issue, but most such studies still have some limitations. First, they tend not to consider the diversity of possible keyword selection strategies, which may vary depending on such factors as which methods are used, how many keywords are extracted and from what part of the patent documentation. There may also be interactions between the factors that should be taken into account in selecting the best strategy. Second, previous studies have seldom examined the process of developing a keyword vector by transforming a keyword set into an analyzable form. Hence, in developing keyword vectors, careful consideration must be given to how the vectors are standardized for an analysis.

3. Criteria of keyword selection and processing strategy for text mining based patent analysis

This paper proposes an effective keyword strategy for text-mining based patent analyses through an evaluation and verification of the keyword selection and processing methods that are most often used in existing studies. For this purpose, the following four factors are addressed: the most appropriate elements of the patent documentation for keyword extraction; keyword selection methods, the number of keywords to be selected; and standardization methods for constructing keyword vectors. Fig. 1 shows the overall process of applying text-mining to patent analysis and the factors that can affect the process.

The text-mining approach is used for transforming unstructured patent data into structured data form. In this step, researchers need to decide which elements of the patent documents they intend to extract keywords from, because they consist of multiple parts, each of which has distinguishing features according to its purpose, and so is likely to yield a different set of keywords. Once this decision has been made, a keyword selection method should be determined, because the criteria for selecting keywords vary. The number of keywords to be selected is also a significant factor – a large number of keywords can be very 'noisy', while too few may be insufficient to represent the overall patent documents. Once the keywords have been selected, they are used to construct a term-vector for each document, which is a list of the document's keywords and their occurrences in that document. Here, how to define and measure components in the term-vector needs to be considered: that is, the data need to be standardized to make the structured result consistent and clear. This factor is worth attention, as the results of analyses using the term-vector can be significantly affected by the standardization methods adopted.

3.1. Elements of patent documentation for keyword extraction

Patent documents are divided into multiple elements, including title, abstract, claims, and description, and because their purposes differ, their sentence structures and vocabulary also differ from each other. First, the title and abstract use distinctive and significantly differentiated words to express the relevant technologies properly, but are short and lack specific details about them. On the other hand, the claims are fuller, and describe the associated technical features explicitly so as to provide full legal protection, which is vital for patents. But, although they are simple and clear, claims are usually written in legal terms, so keywords extracted from these elements may be abstract and lack detailed description of the appearance of the inventions to be patented and the actual functioning of their technologies. Finally, the description contains specific details about the inventions, including the background, summary, and brief description of the inventions, and details of their intended purposes and usages. So descriptions have a lot more content than the other patent document elements, and describe the ideas, methods and processes of the technologies. This can give advantages and disadvantages – while keywords extracted from patent descriptions may include accurate terms to describe technological characteristics of the patent, such terms can also be problematically 'noisy'.

Among these patent documentation elements, most previous studies have extracted keywords from the titles or abstracts (Xie & Miyazaki, 2013). This paper also considers two other sections– the claims and descriptions– and compares their suitability for keyword extraction for patent analysis in innovation studies. Here, it should be noted that the full-text is not considered in this research. Most previous studies on text-mining applications to patents data have adopted only a part of patent documents for keyword extraction possibility because of the following reasons. First, more consistent keywords are expected to be extracted when only a limited part of patent documents are used as different parts of patent documents have different sentence structures and vocabulary. Second, more effective analysis is feasible with a limited part of patent documents as it generally requires a great effort to preprocessing and analyzing the full-text.

3.2. Keyword selection methods

The criteria for selecting keywords from a document may also vary. For example, the words appearing most frequently in particular documents can be regarded as critical, or words that match well with the main document themes are often assumed to be important. In general, while words that appear frequently in patent documents are likely to be representative keywords, those appearing too frequently in such documents are also likely to be general

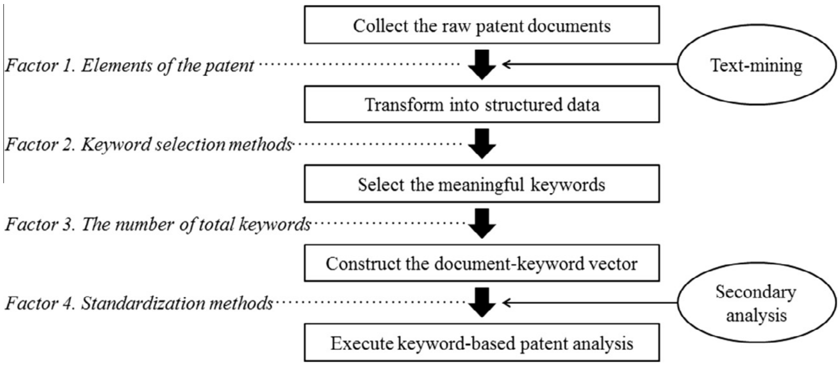

Collect the raw patent documents
Factor 1. Elements of the patent
Text-mining
Transform into structured data
Factor 2. Keyword selection methods
Select the meaningful keywords
Factor 3. The number of total keywords
Construct the document-keyword vector
Factor 4. Standardization methods
Secondary analysis
Execute keyword-based patent analysis

Fig. 1. General process of keyword-based patent analysis and its four factors.

words that are common in all documents, rather than representative words which allow specific patents to be identified.

In existing studies, four major methods – frequency, TF–IDF, variance, and weight-based methods – have been utilized to determine the characteristics of patent documents. Frequency-based keyword extraction targets keywords that appear most frequently in a document (Yoon & Park, 2004), which are selected as keywords because they are emphasized in the document, and are also likely to be highly correlated with other significant terms: indeed, many studies have targeted keywords with a high frequency of appearance (Rokaya, Atlam, Fuketa, Dorji, & Aoe, 2008). In the TF–IDF-based method (the acronym stands for Term Frequency–Inverse Document Frequency), weight values are calculated with multiplying Term Frequency (TF) and Inverse Document Frequency (IDF). TF means the total number of times a word appears in the documents. IDF means an inverse number that the frequency of the word in the corpus. Thus TF–IDF method can describe important keywords which are closely related to representative technology while avoiding general term in the corpus (Usui, Palmes, Nagata, Taniguchi, & Ueda, 2007). The variance-based method selects and uses keywords appearing in documents above a certain level, based on a calculation of the variations in the number of appearances. For instance, Lee et al. (2009) attempted to find new technological areas through a keyword-based patent map approach, and conducted variance and frequency comparison analyses, assuming that extracted keywords with higher variance showed the patent's technical characteristics well. A high variance indicates that a keyword has a high frequency of appearance in specific documents and a low frequency of appearance in other documents, which means it identifies the technical features of a patent document more specifically. Finally, the weight-based method has the valuable advantage that selecting keywords via a weighting calculation can represent overall documents without considering their total lengths – as opposed to the frequency-based approach, which is unsuitable for comparing documents of different lengths. Tseng et al. (2007) aimed at developing automated patent classification and analysis methods by applying text-mining to patent documents, and trying to enhance the efficiency of this process. In particular, the authors extracted keywords by imposing a weight which is based on not only the number of words appearing, but also the word relationships between headings and bodies of the patent document. In addition, Edmundson's study and Slaton and Buckley's study on keyword extraction proposed an automatic extraction technique by combining basic methods in which keywords were selected by adding the word appearance frequency to the weight (Edmundson, 1969; Salton & Buckley, 1988). Of these four types of methods, this study focuses on frequency, variance, and TF–IDF, which have been used frequently in related research.

3.3. The number of keywords extracted

Text-mining based patent analyses use a quantitative approach, rather than using experts to select keywords, or dictionaries of technical terms, and thus may have more 'noisy' terms. Nevertheless, text-mining based approaches are fast and easy to use, so they can reduce the time and cost resources needed to extract keywords. From this viewpoint, the number of keywords selected is a critical factor in text-mining based patent analyses: if too many are selected, it is likely that too many general words will be included as well as the important ones; if too few, it is likely that keywords that appear only in specific documents get extracted, making it difficult to show the overall documents' characteristics effectively. The number of keywords used in the existing research generally falls between 30 and 100 items (Li et al., 2009; Tseng, 2005) – but it is hard to find such studies which focus on this factor. For example, Lee, Lee, Seoul, and Park (2008) identified 242, 259 and 563 keywords from 257, 552 and 762 patents respectively but used only 29 meaningful ones for further analysis to develop a keyword-based technology roadmap. A similar number was used in the work by Lee et al. (2009), where 39 keywords were selected from 141 PDA-related patents and were used to develop a keyword-based patent map. Whereas, less than 30 keywords have also been adopted for analysis. Seol et al. (2011) applied text mining to extract keywords that represent the attributes of products from patent documents and then selected only 10 keywords to measure the technological strength by products. Therefore, we first explore what range of the number is suitable for keyword-based patent analysis: for this purpose, this research identified 0 to 150 keywords as an explorable range.

3.4. Term vector standardization methods

A term vector is an algebraic expression that describes the relationship between text words and documents, and is commonly used as a dataset for text-mining based analysis. In this method, each dimension of the vector corresponds to an individual term, which can be a single word or keyword or sometimes a longer phrase. If a specific document includes a specific term, the vector value of that term should be more than 0.

In using term vectors, which are based on the keywords selected for text-mining analysis, it is imperative to consider how they should be standardized. Many researchers have paid attention to automatic patent classification by applying an SVM (support vector machine), NB (Naïve Bayes), or k-NN (k-Nearest Neighbor) algorithm (or other clustering methods) to term vectors (Cong & Tong, 2008; Loh, He, & Shen, 2006), and have commonly conducted, standardization on the use of datasets as a preprocessing process before implementing these methods. A term frequency may be high possibly because of the two reasons: firstly, keywords can appear frequently in a document being analyzed on account of its importance; and secondly, the document may be long enough to allow for such high frequency. The standardization process attempts to retain the former reason, but eliminate the latter. Standardization affects the results of cluster analyses, but previous studies have been subject to the limitation of rarely taking into account the fact that standardization methods may affect analysis performance (Milligan & Cooper, 1988). Similarly, as a number of patent analyses applying text-mining use a term vector for secondary analysis, standardization strategies needs to be investigated in full.

The method described in this paper, which examines the representativeness of a keyword set using k-means clustering, applies three commonly used standardization methods for document analysis. The first method carries out no standardization of the variables; the second standardizes them by processing them into a range of 0–1; the third uses a Boolean expression to identify whether a corresponding word appears in a specific document or not.

4. Research method

4.1. Overall research process

The four factors noted above can all affect the results of keyword-based patent analysis, so a strategy that addresses them in combination when selecting and processing a keyword set from patent documents is required. To address this need, this research adopts three methods – orthogonal array, k-means clustering, and entropy value – as its main research methods in its aim to build some 'best guidelines' for keyword-based patent analysis. Fig. 2 represents the overall research framework.

First, a total of 500 patent documents from five USPC (US Patent Classification) classes were collected. Using these patent documents as a raw data set, basic keywords were extracted by TextAnalyst2.1 and various keyword vectors derived by applying keyword selection and processing strategies. Next, as an exploratory study, the keywords sets extracted by different keyword selection strategies were compared to see whether they produced different keyword sets or not: if so, the urgent need for this research would be illustrated.

After that, an orthogonal array was applied to design experiments, each of which represents a combination of the four keyword-based patent analysis factors. In each experiment, k-means clustering and entropy value were used to calculate the performance of the particular keyword selection strategy. The results of k-means clustering would be acceptable if term vectors developed under the strategies were good enough to cluster patent documents from the same USPC class. If clustering analysis based on their term vectors grouped all patents in the same USPC class precisely together in five clusters, the entropy value would be zero: on the other hand, if the analysis scattered patents from the same USPC classes across different clusters, the value should be closer to 1, indicating a high degree of disorder. Finally, ANOVA (Analysis of Variance) was conducted to investigate the effects of the four factors noted above, the result of which indicated the most effective keyword selection and processing strategy as a guideline for text-mining based patent analysis.

4.2. Detailed procedures

4.2.1. Patent data collections

The data used for this research are patent documents from five USPCs. As this classification system uses experts to classify patent documents according to their contents, patent documents in the same class can be expected to have similar keywords, which will differ from those in documents from other classes. The technical areas to be studied for this research, and the classes belonging to related technical areas, were therefore assigned, and the corresponding patent documents were collected.

Alternative energy-related technology areas were adopted as the focus of this research: five relevant USPC classes were selected as shown in Table 1 and 100 patent documents for each class were randomly selected, giving an overall total of 500 documents. The alternative energy field has recently been expected to undergo high technological growth, and its related technology development is very active; so technology trend analyses using patent data in this particular area are being actively conducted. In addition, as this area has a different knowledge background and method approach from other technology groups, but is common in terms of the energy-related technologies involved, it is characterized as satisfying both generality and specificity for keywords extracted through a text-mining approach.

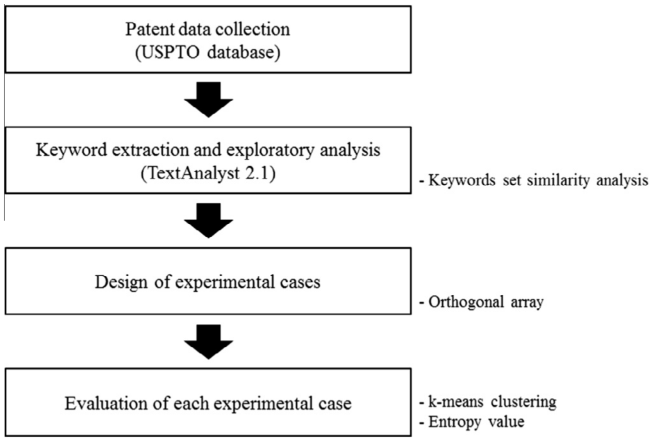

Patent data collection
(USPTO database)
Keyword extraction and exploratory analysis
(TextAnalyst 2.1)
- Keywords set similarity analysis
Design of experimental cases
- Orthogonal array
Evaluation of each experimental case
- k-means clustering
- Entropy value

Fig. 2. Overall research framework.

Table 1
Alternative energy production USPC codes.

Classes	USPC code
Class 1	Gasification (48/197R)
Class 2	Genetically Engineered organism (435/252.3)
Class 3	Solar cell (438/57)
Class 4	For passive space heating (52/173.3)
Class 5	Wind (290/44)

4.2.2. Keyword extraction and exploratory analysis

TextAnalyst 2.1 was used to extract keywords from the text documents. The program operates with a neural network and a semantic analysis. Because semantic analysis cannot be conducted as a self-learning process, a neural network is automatically built for the imported texts, and then a semantic structure is created based on that network. After TextAnalyst 2.1 extracted the keywords, an exploratory similarity analysis was conducted to see how many of the same keywords were extracted by different approaches. If the similarity varies depending on the number of selected keywords, it will not only prove the necessity of the current research, but it will also offer suggestions as to an appropriate level of the number of keywords. If c_1 is the keyword set for the comparison group 1, c_2 the keyword set for comparison group 2, and N is the number of keywords, the similarity index can be defined by:

$$\textit{Similarity index} = \frac{n(c_1 \cap c_2)}{N} \times 100(N = \mathrm{n}(c_1) = n(c_2)) \tag{1}$$

Since the similarity index used for this research is the percentage of the keywords that are common to both groups 1 and 2, and if all keywords extracted from comparative groups are identical, the index value would be 100: the fewer keywords are in common between the two groups, the closer the index is to 0. Therefore, using the similarity index, we can easily compare the share of keywords different approaches (e.g. using different parts of patent information or using different keyword selection methods) produced in common with a same scale ranging from 0 to 100.

4.2.3. Experiment design

An orthogonal array is widely used for experimental design due to its distinguishing advantages (Hedayat, Sloane, & Stufken, 1999; Kuhfeld & Suen, 2005; Wang, Tang, & Zhang, 2011). Users can design experiments systematically through the orthogonal array with no statistical background. Also, they can place various factors into an experiment without increasing the experiment's size. In this research, we adopted the orthogonal array to take these advantages, especially the second one.

Interaction effects with more than three factors can be taken as meaningless (or very small) and so can be assumed to be irrelevant. The purposes of this research can only be achieved by observing all secondary interaction effects for three residual factors, i.e., excluding factor D which is not directly connected to a patent, but is introduced by term vector standardization methods. Thus the observations of this research were limited the main effects of the four factors previously mentioned, and their three secondary interaction effects. So a total of seven effects were identified as necessary to determine the outcomes effects of this research, as shown in Table 2.

Table 2
Experimental case factors.

Factor	Criterion
A	Elements of paten documentation for keyword extraction
B	Keyword selection methods
C	The number of keywords extracted
D	Term vector standardization methods
A*B	Keyword selection methods under a specific element of patent
A*C	Elements of the patent under a specific number of keywords extracted
B*C	Keyword selection methods under a specific number of keywords extracted

4.2.4. Evaluation of each case

To evaluate the performance of the keyword selection and processing strategy in each experimental case, this study used k-means clustering and measured the resulting entropy values. Clustering analysis is one of the most widely used methods in various types of research, and is often found in automatic patent classification studies (Daim et al., 2006). If an extracted keyword can represent a patent's overall documents, the quality of the clustering result will be high. Entropy is a barometer, originating from information theory, used to measure the homogeneity within clusters, and acts as a classifier for a series of datasets, and calculating and digitizing existing class variances for documents based on the clustering results. If a cluster contains only one identical class document, the entropy value will be 0. If the cluster contains several classes of documents, the entropy value increases. If there are N original document clusters for n documents, a clustering analysis using the extracted keyword set produces M new clusters (as shown in Fig. 3) and the entropy of each new cluster C_j can be shown as:

$$e(C_j) = \sum_{i}^{N}\left(-\frac{X_{ij}}{N}\log_N\frac{X_{ij}}{N}\right) \quad (i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, M) \tag{2}$$

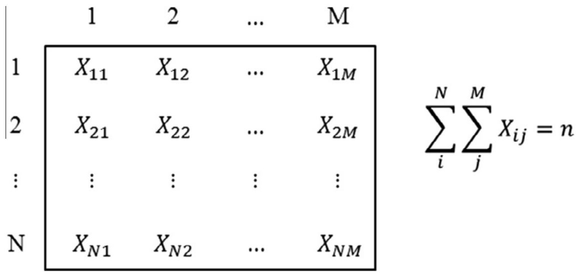
In addition, since the number of patent documents assigned to each new cluster is not constant, it is necessary to calculate the weight for each cluster using:

$$w(C_j) = \frac{1}{n}\sum_{i}^{N}X_{ij} \tag{3}$$

Thus, if the selected keyword set represents the whole of the patent documentation well, the entropy value for whole patents as given by:

$$e = \sum_{j}^{M}w(C_j)e(C_j) \tag{4}$$

will be close to 0; and the less that set represents the patent documents, the nearer the cluster's entropy value will be to 1.



$$\begin{array}{c|cccc} & 1 & 2 & \cdots & M \\ \hline 1 & X_{11} & X_{12} & \cdots & X_{1M} \\ 2 & X_{21} & X_{22} & \cdots & X_{2M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ N & X_{N1} & X_{N2} & \cdots & X_{NM} \end{array} \qquad \sum_{i}^{N}\sum_{j}^{M}X_{ij} = n$$

Fig. 3. Cluster matrix for entropy calculation.

Fig. 4 shows the process of assessing the performance of each keyword selection and processing strategy.

5. Analysis results

5.1. Similarity analysis of keyword set

Of the four factors, three (excluding standardization) are closely related to keyword selection strategy from patent documents. Therefore, before investigating the performance of the derived keyword set through clustering, this study needs to examine how many of the same words from the keyword set are extracted under those three factors. Homogeneity between keyword sets is measured based by the similarity index given in Eq. (1). This study calculates similarity indexes for two observed cases: first how the similarity values change depending on the elements of the patent documentation involved and the total number of keywords extracted by the different keyword selection methods, i.e. frequency, variance and TF–IDF; and second, how they change depending on the keyword selection method and the total number of keywords extracted from different patent elements, i.e. abstract, claims, and description.

Fig. 5 illustrates the first case, and shows that, while the similarity between the selected keyword set in the abstracts and claims is high, similarities between the abstracts and descriptions, and between the claims and descriptions, are low. And the similarity values tend to converge as the total numbers of keywords increases from lower to higher numbers.

In the second case, Fig. 6 shows the similarities when different keyword selection methods are used to select keyword sets for each type of patent document. The figure shows that although the claims element consists mainly of legal terms, the keywords extracted by text-mining from the abstract and the claims might be very similar: on the other hand, the words in the description are somewhat different from those in other patent elements.

Table 3 shows the results of an ANOVA conducted to examine whether there is any significant difference in the keyword set obtained when using different keyword selection methods on the same patent document elements. The analysis results show that the similarity for the keyword set has a statistically significant difference between keyword selection methods when used on the same patent document elements, and between patent document elements when subject to the same keyword selection method. Detailed results of ANOVA are described as Tables A.1 and A.2 in Appendix A.

The two kinds of similarity analysis and the ANOVA result above show that the keywords extracted differ according to the keyword selection methods, the elements of the patent documents involved, and the numbers of keywords extracted, indicating the need for a guideline keyword selection strategy to be established for a patent analysis using text-mining. In addition, although this study did not determine the appropriate levels of the number of keywords factor at Section 3, the levels of the factor can be determined through the similarity analysis in this section. In this paper, the levels are determined from changes of similarity pattern from Figs. 5 and 6. First, similarity values increase as the numbers of keywords goes up towards 30. This phenomenon is almost same for the all similarity graphs. Second, some similarity patterns go up and others go down in a range between 30 and 70. Although it is hard to describe why similarity indices are irregular in this range, but, pattern changes between the range of 0–30 and the range of 30–70 shows opposite direction in most similarity graphs. Third, similarity indexes take little changes when the number of keywords is between 70 and 130. In this range, most graphs shows converging pattern against other ranges. Although these patterns

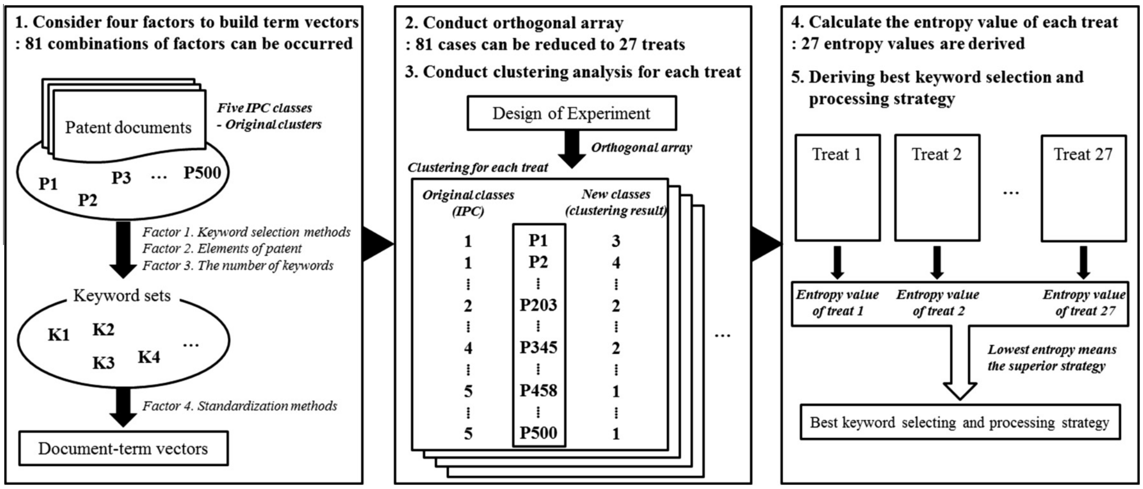

1. Consider four factors to build term vectors
: 81 combinations of factors can be occurred
Patent documents
Five IPC classes
- Original clusters
P1
P2
P3
…
P500
Factor 1. Keyword selection methods
Factor 2. Elements of patent
Factor 3. The number of keywords
Keyword sets
K1
K2
K3
K4
…
Factor 4. Standardization methods
Document-term vectors
2. Conduct orthogonal array
: 81 cases can be reduced to 27 treats
3. Conduct clustering analysis for each treat
Design of Experiment
Orthogonal array
Clustering for each treat
Original classes (IPC)
New classes (clustering result)
1 P1 3
1 P2 4
2 P203 2
4 P345 2
5 P458 1
5 P500 1
…
4. Calculate the entropy value of each treat
: 27 entropy values are derived
5. Deriving best keyword selection and processing strategy
Treat 1
Treat 2
…
Treat 27
Entropy value of treat 1
Entropy value of treat 2
Entropy value of treat 27
Lowest entropy means the superior strategy
Best keyword selecting and processing strategy

Fig. 4. Process for assessing the performance of each keyword strategy.

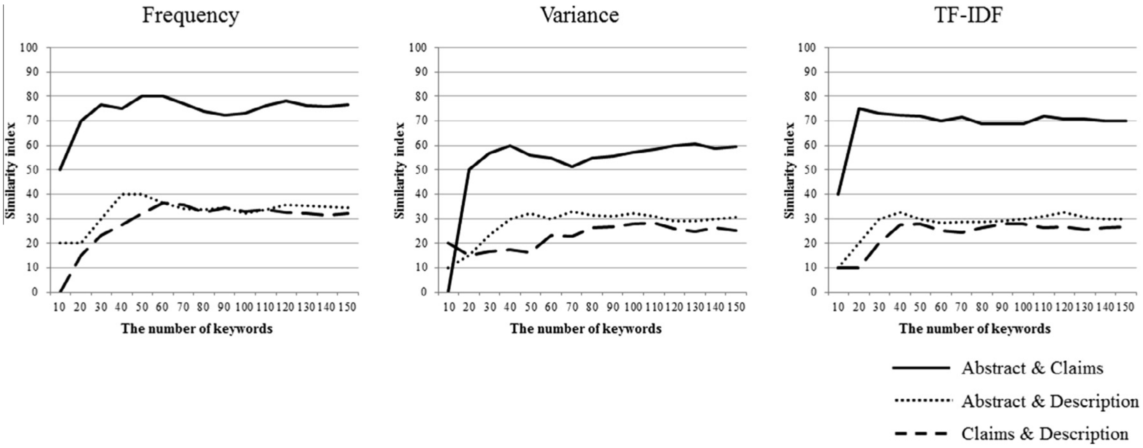


Fig. 5. The same selection methods applied to different patent document elements.

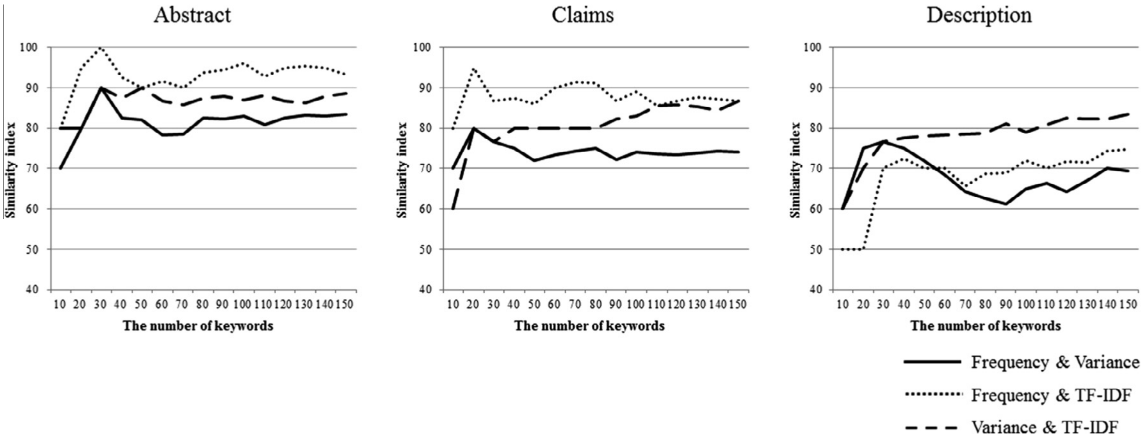


Fig. 6. Different keyword selection methods applied to the same document elements.

Table 3
ANOVA similarity analysis results.

Keyword selection method	Source of variation	*p*-Value	Elements of the patent	Source of variation	*p*-Value
Frequency	Elements of the patent	.000	Abstract	Keyword selection methods	.000
Variance		.000	Claims		.000
TF–IDF		.000	Description		.000

can only be observed through graphs, most patent analysis studies use around 100 keywords, and a change in keyword similarity indicates that there is a difference in the term vector. Thus this research uses 30, 70, and 130 as the levels of the number of keywords extracted.

5.2. Orthogonal array for experiment design

This study identified the necessity of this research through the similarity analysis, and determined the appropriate levels of the number of keywords above. Thus, in all, four factors have to be examined, and there are three levels for each factor, as Table 4 shows. As the total number of experiments based on all of these criteria is 3^4 (i.e., 81) a lot of effort is needed to carry out the k-means clustering and calculating the entropy values: so this study proposes a method of reducing the number of experiments required. In addition, combinations of the factors need to be considered to determine the best strategies for keyword selection and processing, since the performance of the keyword set can be affected not only by the factors, but also the interactions between them. This study takes these problems into account so that the orthogonal array can be used properly.

Table 5 shows the results of the ANOVA analysis for the major and interaction effects that are important to keyword selection and processing for a keyword-based patent analysis. These results show that the patent document elements, the keyword selection methods, and the number of keywords are not statistically significant, and that the interaction effects are also statistically insignificant. In contrast, standardization methods are shown to be statistically significant, which means the clustering result is affected by the standardization method chosen: so this needs to be taken into account when the term vector is used as the dataset for additional analysis.

Although factors other than standardization do not affect the performance of the keyword selection and processing strategy, the entropy value from the experimental results varies greatly, from 0.18 to 0.80. Although the factors used to select and process the keywords are not statistically significant, as the differences in entropy values are likely to affect the reliability and the results of a secondary analysis in a patent study using text-mining, identifying the best strategy for keyword selection and processing is still very important: this is drawn up in the next section.

5.3. Best strategy for selecting keywords for a patent analysis

As the patent document elements, keyword selection methods, and total number of keywords are not statistically significant, this paper does not argue that these factors are important for the best keyword selection and processing. Rather this study contends that the combination that most reduces the entropy value is the best strategy for a keyword selection and processing strategy, and so proposes to identify the best strategy for patent analysis based on the line chart shown as Fig. 7. Detailed entropy values of each level of factors are described as Table A.3 in Appendix A.

Because an entropy value close to 0 indicates better performance, it can be argued that the best keyword selection and processing strategy consists of those factors whose estimated marginal mean values are low on their individual charts. As the factors' interactions are proven to be not statistically significant, individual line charts can show which level of each factor is the best for keyword selection and processing strategy.

Table 4
Factor levels.

Factor	Levels
Elements of the patent	Abstract, claims, description
Keyword selection methods	Frequency, variance, TF-IDF
The number of keywords extracted	30, 70, 130
Term vector standardization methods	None, range (0–1), Boolean expression

First, for the patent documents (see line chart (a)), the abstract and claims have little difference in entropy value and the value for the description is considerably greater, since that element may have more text than the abstract and claims, and the words in the description may be noisy.

Second, as the frequency factor in the keyword selection method only considers the numbers of words' appearances in an entire document, this factor less likely allow for the identification of specific patents (see chart (b)). On the other hand, the TF–IDF method has the lowest entropy value because it is believed to be influenced from the features of both frequency and variance. It is interesting that the variance-based keyword selection method shows a lower entropy value than the frequency-based method, which means the variance method can be a useful way to select keywords from patent documents, despite the fact that many studies on keyword-based patent analysis have used the frequency method as the key criterion for keyword selection. Also, there is little difference in entropy value between the variance and TF–IDF methods, so researchers can decide between them according to their relative accuracy and the amount of effort they involve.

Third, for the number of keywords (see chart (c)), our results show that the entropy value decreases as the number of keywords increases, which is probably caused by the fact that the numbers of keywords increase with the information in the whole document. However, the entropy gap is greater when the number of keywords changes from 30 to 70 than from 70 to 130 keywords, which means that, as the number of keywords increases, the amount of information from the keywords set converges to the critical point (as in the case of the similarity analysis described in this study): in other words, when the number of keywords increases above 130, the entropy value seems to be only marginally reduced.

Finally, for the standardization method (chart (d)), a Boolean expression has an extremely low entropy value, which means that when already-selected keywords are used to build a term vector, information on whether words have appeared in certain documents, rather than the frequency of their appearance, can be the most meaningful output.

So using the TF–IDF method to extract 130 keywords from an abstract appears to be the best keyword selection strategy, and using a Boolean expression will provide the best results when a cluster analysis is conducted with the term vector as the dataset.

6. Discussion

Text-mining based patent analysis is different from other text-mining applications, because, although patent documents may be described in natural language – like other kinds of text documents – they also contain formal words about specific technologies and are set out in structured formats. These specific characteristics have led a number of researchers to apply text-mining techniques to investigate patents, and these previous studies have revealed some factors that need to be considered when using this process for patent analysis. This research has focused on four of these factors and suggested a best keyword selection and processing strategy that could be adopted. But, although the final result of this study has both theoretical and practical implications, there still exist matters that must be discussed.

First, the description element of patents' documentation can be more useful than the abstract and claims when using a keyword-based approach to compare patent documents in the same technology field. In fact, it is the abstract and claims that are most widely

Table 5
ANOVA keyword selection treatment results.

Source of variation	SS	d.f	MS	F	*p*-Value
Corrected model	.981	20	.049	6.922	.012
Intercept	7.307	1	7.307	1031.062	.000
Elements of the patent	.001	2	.001	.097	.909
Keyword selection methods	.014	2	.007	1.012	.418
The number of keywords extracted	.010	2	.005	.698	.534
Term vector standardization methods	.860	2	.430	60.673	.000
Elements of the patent X Keyword selection methods	.004	4	.001	.128	.967
Elements of the patent X The number of keywords extracted	.007	4	.002	.259	.894
Keyword selection methods X The number of keywords extracted	.085	4	.021	2.984	.112
Error	.043	6	.007		
Total	8.331	27			
Corrected total	1.024	26			

Note: SS: sum of square/d.f: degree of freedom/MS: mean square/F: F statistic.

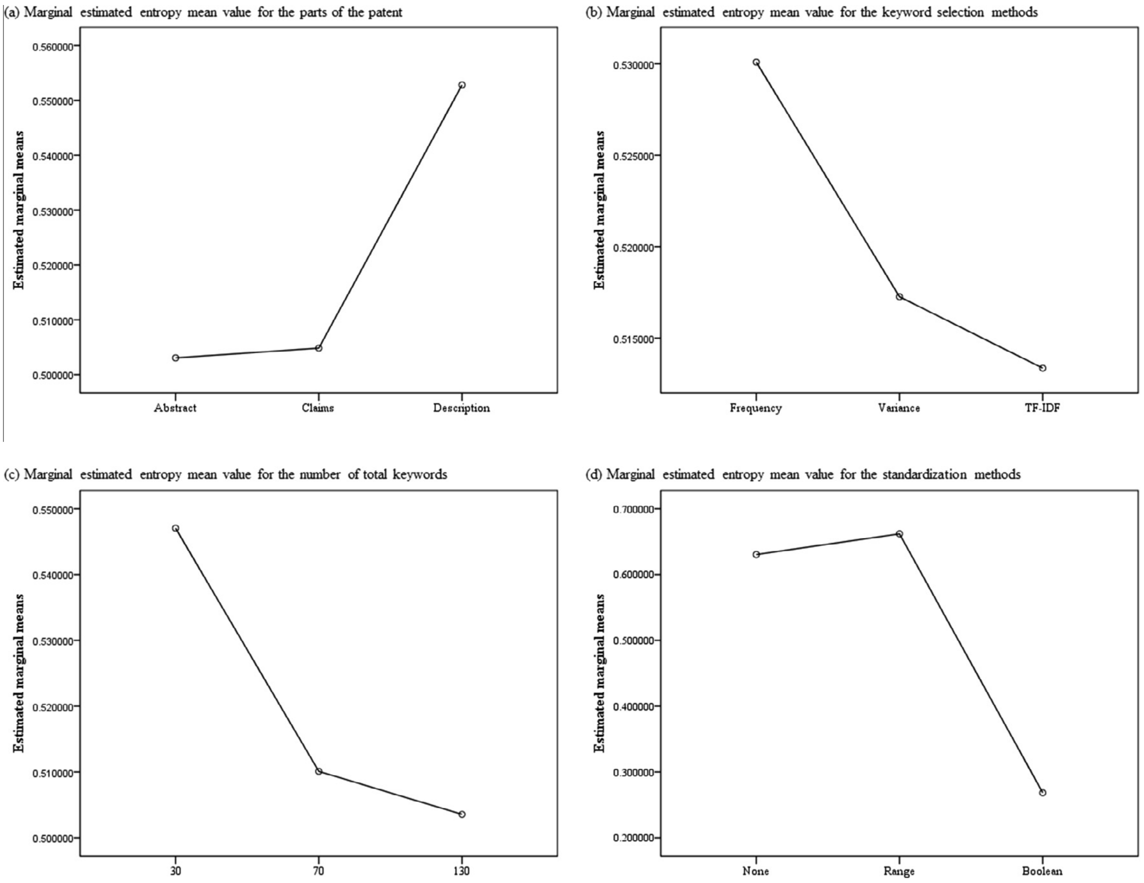


Fig. 7. Entropy value line charts.

used for patent analysis: researchers generally regard the description as having too much information to yield meaningful insights. However, this noisy feature can be helpful when researchers want to compare patents in same technological field in detail. As Fig. 4 shows, word composition similarity is low when the description is adopted for keyword approaches. This study chose five different classification classes in the alternative energy production field in this study, but the similarity between the patents' abstracts and claims is high, which means researchers cannot precisely investigate patents in these same technology fields by utilizing keywords extracted from these elements. In this case, keywords from the patent description can be more meaningful, and can be a valuable way to build up a strong patent portfolio by comparing competitor's patents at the same technology field.

Second, the keyword selection method can be selected differently according to a purpose of the patent analysis. Although this study shows TF–IDF to be the best overall method for a text-mining approach, its results show that word component similarities do not differ much across keyword selection methods. Especially, the similarity between the frequency and TF–IDF methods as applied to patent abstract and claims elements is very high – Fig. 5 shows their similarity to be very close to 90%. At the same time, the similarity between the variance and TF–IDF methods is relatively high for the description elements. These results may be accounted for by the different characteristics of the patent elements: the frequency and TF–IDF methods produce similar keywords for the abstract and claims elements owing to their conciseness, while the variance and TF–IDF methods may yield similar keywords when applied to description elements, because of their noisy nature. So TF–IDF is not necessarily the only route for good keyword selection strategy - researchers can choose an appropriate method depending on their research purpose or other conditions (e.g. available time and efforts for analysis).

Third, it might be less important to deliberate whether each keyword from patent documents extracted is meaningful or not when they are extracted for the purpose of constructing a document-keyword vector. In this study, word composition similarity varies considerably according to the factors examined. Table 3 shows that the keyword selection methods and the elements of the patent documents significantly affect word composition similarity. On the other hand, those factors do not affect the performance of clustering results (as shown in Table A.1 – see the Appendix A). This means that, even if different keyword selection strategies produce disparate groups of extracted words, affect patent document classification will not be affected. Researchers should focus on which set of words can best represent the patent documents overall, rather than focusing on which words seem most important and essential in individual documents. Patent documents describe a specific technological theme: while selected words may differ depending on the keyword selection strategy adopted, all of them will be more or less related to the patent's narrow technological field.

Fourth, this study adopted the entropy index to measure the performance of keyword selection and processing strategy. In this research, a keyword vector was developed for each of the 500 patents, patents randomly selected from five distinct USPC classes, and, based on which clustering analysis was conducted to classify the patents into five homogeneous groups. Then, the original USPC classes were compared with the clustering results using the entropy value on the assumption that the USPC system is a reliable technology classification system and so patent documents belonging to the same class are similar to each other in their contents. Though the entropy index is a meaningful performance index, which measures the extent to which clustered labels match externally supplied class labels, it hardly measures the goodness of a clustering structure and provides only limited information about how close patents clustered under the same labels are. There are other measures worth considering depending on the target for performance or quality. For example, a text-mining approach is applicable not only to group patent documents but also to other areas such as auto-patent retrieval, auto-patent classification, and technology ontology development. In these cases, a more precise index should be designed to measure the performance of keyword selection and processing strategy at the individual (patent) level, not the group (USPC class) level. Sometimes, relying on experts' opinion may also be effective in evaluating the performance. In addition, the goodness of a clustering result can be affected by clustering methods. In this research, we adopted a k-means clustering but other clustering algorithms may produce different results. Therefore, researchers should carefully design a performance index depending on the context. Examining the overall performance based on several indices simultaneously will also help improve the reliability of analysis results.

Fifth, computational resources were not considered in this research because only 500 patents were used to evaluate the various types of keyword selection and processing strategy. However, when the number of patent documents increases and the amount of computational resources is limited, the resources spent on implementing the strategy become significant and should be taken into account. The computational resources of each level within the four factors are different. Firstly, as to the elements of the patents, the required computational resources increase with the length of texts. And thus *description* will require the most computational resources, followed by *claims* and *abstract*. Actually, the total file size of the 500 patents used in this research was 342 KB for *abstract*, 3,071 KB for *claims* (almost nine times larger than *abstract*), and 33,184 KB for *description* (almost 10 times larger than *claims* and 93 times larger than *abstract*). Secondly, with regard to the keyword selection methods, the required computational resources depend on their computational complexity. Hence, the resources for *TF–IDF* would be the highest, followed by *variance* and *frequency*. Thirdly, relating to the number of keywords extracted, the required computational resources grow with the number of keywords. The number of keywords corresponds to the number of data features. More computational effort will be needed when more features are used. Accordingly, the most resources are required for *150 keywords* while the least are needed for *one keyword*. Finally in the fourth factor of term vector standardization methods, the required computational resources for *range* are expected to be larger than *Boolean expression*. Therefore, the best keyword selection and processing strategy may change according to the computational resources allowed, which is particularly significant in analyzing a large number of patent documents.

These five discussions will be give greater insights as text-mining becomes more commonly used in the future for firms' Intellectual Property (IP) management, and investigating the factors involved may help reduce the costs and efforts required to analyze IP documents (including patents). Thus, understanding the discussion and limitations of this research can be valuable for the further study.

7. Conclusion and future research directions

The purpose of this study is to suggest guidelines for selecting and processing keyword sets for using text-mining in patent analysis. Four different factors have been considered, and the performance of a keyword set evaluated based on clustering analysis and entropy values. The patent document elements, keyword selection methods, and the total number of keywords were proven to be statistically insignificant, which can be interpreted as meaning that existing studies on keyword-based patent analyses conducted without a standard criterion for selecting keywords all yield equally reliable insights. In other words, the differences in studies on keyword selection from an abstract based on the frequency of keywords, or studies based on the variance of terms in the claims, are not critically enough to affect the keyword-based patent analysis results. But the study found that selecting 130 words from an abstract based on a TF–IDF and Boolean expression appeared to represent the best keyword selection and processing strategy, and thus most suitable for patent research. Although many studies into keyword-based patent analyses use TF–IDF to extract hundreds keywords from the patent abstract, very few studies have investigated whether such a method is the most effective.

This study's results make some contributions in both theoretical and practical perspectives. First, theoretically, it could establish a basis for future text-mining applied patent analysis. There are many factors or criteria – such as the number of keywords or keyword selection methods – to be considered before designing keyword-based patent research. But it has proved hard to find studies that consider these kinds of factors, which is why this study has investigated a variety of factors related to text-mining based patent analysis. As a result, it contributes to improving the reliability of existing keyword-based patent analysis studies by identifying the guidelines for choosing the best keyword selection and processing strategy, as well as suggesting the best keyword selection strategy on the basis of that which has the lowest entropy value. Second, this study offers practical guidelines for IP management, which has evolved recently with development of data analysis techniques. As the importance of building strong IP portfolios increases, keyword-based patent maps have been widely applied to identify the technical limitations of a firm's technology, or its potential market opportunities. Nevertheless, extraction of meaningful keywords from patent documents is not easily carried out in practice. Conducting keyword extraction via the conventional approach, which uses groups of experts to draw patent maps, demands considerable resources. But, in applying text-mining approach, managers may find it hard to extract meaningful keyword lists from patent documents owing to their insufficient data mining knowledge, or may not be convinced that using a mathematical approach will extract sufficient keywords to draw meaningful patent maps. In this case, the best keyword selection and processing strategy proposed in this paper may offer a simple and a helpful alternative for drawing keyword-based patent maps.

Until now, little effort has been made to investigate an effective strategy for keyword selection and processing in the context of keyword-based patent analysis. Although recent studies have suggested advanced methods to extract and analyze keywords from patent documents, reflecting the growing interest in keyword-based patent analysis (e.g. Fu et al. (2013), Fu, Chan, et al. (2013), Jeong and Kim (2014), Murphy et al. (2014), Park, Kim, Choi, and Yoon (2013), Park, Ree, and Kim (2013)), few of them have addressed the rationales for the use of a particular keyword selection and processing method, relating to the four factors in this research – elements of patent documentation for keyword extraction, keyword selection methods, the number of keywords extracted, and term vector standardization methods. Quite frequently, a part of patent documents for keyword extraction was chosen without giving a reason for the choice and the number of keywords used for further analysis was not even mentioned. Recognizing the research need, a few researchers dealt with the relevant issues. For example, Xie and Miyazaki (2013) evaluated the representativeness of a keyword set constructed from different part of patents. However, further analysis is needed considering the combination of various factors affecting the performance of keyword selection and processing strategy. Our study aims to contribute to fill this research gap, being supplementary to the existing studies.

Despite those insightful contributions, some limitations still remain. First, only k-means method is used to examine the performance of keyword selection and processing strategy. Although the method is one of representative clustering analysis, clustering results might be changed depending on the clustering method. A second limitation concerns the standardization method for term vectors, which this study regards as an important factor for analyses using a term vectors as their datasets. Moreover, the performance of a standardization method also varies depending on the analysis techniques applied, and so may need to be re-considered for different analysis methods. Data standardization methods are diverse, so the three levels of this study cannot be used as representative of all standardization methods.

Thus, future research directions may include diversifying the keyword selection and processing strategy and elaborating its performance indices. Firstly, more factors affecting the performance need to be examined. For example, the full-text of a patent, which was not considered in this research, might be a valuable source for the keyword-based patent analysis, if well analyzed, and thus can be considered. More data standardization methods need to be investigated. Also, the use a subject-action-object (SAO) analysis to better understand the roles of words can be added to one of the factors. Secondly, more performance indices, in addition to the entropy value, are worth being developed and applied to patent data. The possible options include using internal indices such as the sum of square or designing other external indices relying on experts' opinion. Besides, the balance between the depth of analysis and the resources required for analysis is significant in evaluating the performance. Also, performance targets may vary by the purpose of text-mining applications to patents. And so, more performance indices can be suggested, reflecting the context in which the performance evaluation occurs. Thirdly, in a similar vein, other classifiers or clustering methods are available for clustering patents. Different classifiers or clustering methods may result in different clustering results. Selecting an appropriate classifier or clustering method is critical to ensuring the validity of research, which requires further consideration. Finally, elaborating a keyword extraction method can affect the performance of strategy. A description part of a patent document has a potential to provide more valuable knowledge when more advanced keyword extraction techniques are used. In this research, we relied on a text-mining solution to extract keywords but keyword extraction algorithms vary by solutions. Although the research findings in Fig. 7 indicated that the description part has the highest entropy value possibility due to its noisy words, we can expect more meaningful implications from the description part if those noisy words are removed effectively with the keyword extraction method. A well-structured ontology for technologies, called technology dictionary, can be developed to increase the quality of analysis, where WordNet, which is a semantic network database for English developed at Princeton University is used, or latent semantic analysis is conducted on the relevant patent documents. How to develop and update the technology dictionary is a valuable topic to explore.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIP) (NRF-2013R1A2A2A03016904). This work was also partially supported by the Ajou University research fund.

Appendix A

See Tables A.1–A.3.

Table A.1
ANOVA results of similarity analyses under same keyword selection methods.

Keyword selection methods	Source of variation	SS	d.f	MS	F	Significance
Frequency	Elements of patent	18749.260	2	9374.630	156.809	.000
	Error	2510.920	42	59.784		
	Total	21260.180	44			
Variance	Elements of patent	7782.737	2	3891.368	40.393	.000
	Error	4046.137	42	96.337		
	Total	11828.874	44			
TF–IDF	Elements of patent	18596.009	2	9298.005	204.799	.000
	Error	1906.828	42	45.401		
	Total	20502.837	44			

Note: SS: sum of square/d.f: degree of freedom/MS: mean square/F: F statistic.

Table A.2
ANOVA results of a similarity analysis under the same patent document elements.

Elements of patent	Source of variation	SS	d.f	MS	F	Significance
Abstract	Keyword selection methods	1000.969	2	500.484	33.384	.000
	Error	629.650	42	14.992		
	Total	1630.619	44			
Claims	Keyword selection methods	1409.073	2	704.536	36.880	.000
	Error	802.348	42	19.104		
	Total	2211.421	44			
Description	Keyword selection methods	1010.775	2	505.387	12.570	.000
	Error	1688.581	42	40.204		
	Total	2699.356	44			

Note: SS: sum of square/d.f: degree of freedom/MS: mean square/F: F statistic.

Table A.3
Estimated marginal entropy mean values of line graphs.

Factors	Levels	Estimated marginal means
Elements of the patent	Abstract	.503
	Claims	.505
	Description	.553
Keyword selection methods	Frequency	.530
	Variance	.517
	TF–IDF	.513
The number of keywords extracted	30	.547
	70	.510
	130	.504
Term vector standardization method	None	.630
	Range	.662
	Boolean	.268

References

Basberg, B. L. (1987). Patents and the measurement of technological change: A survey of the literature. *Research Policy, 16*(2), 131–141.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Prentice-Hall Inc.

Chakrabarti, S., Dom, B., Agrawal, R., & Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal, 7*(3), 163–178.

Cheong, H., Chiu, I., Shu, L. H., Stone, R. B., & McAdams, D. A. (2011). Biologically meaningful keywords for functional terms of the functional basis. *Journal of Mechanical Design, 133*(2), 021007.

Choi, J., & Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting and Social Change, 83*, 170–182.

Clifton, C., Cooley, R., & Rennie, J. (2004). TopCat: Aata mining for topic identification in a text corpus. *IEEE Transactions on Knowledge and Data Engineering, 16*(8), 949–964.

Cong, H., & Tong, L. H. (2008). Grouping of TRIZ inventive principles to facilitate automatic patent classification. *Expert Systems with Applications, 34*(1), 788–795.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, 73*(8), 981–1012.

Dou, H., Leveillé, V., Manullang, S., & Dou, J. M. Jr, (2005). Patent analysis for competitive technical intelligence and innovative thinking. *Data Science Journal, 4*, 209–236.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM), 16*(2), 264–285.

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information, 25*(3), 233–242.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1986). The structure-mapping engine. Department of Computer Science, University of Illinois at Urbana-Champaign.

Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003). Automated categorization in the international patent classification. In *ACM SIGIR Forum* (Vol. 37, no. 1, pp. 10–25): ACM.

Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: A practical business case. *World Patent Information, 25*(4), 335–342.

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics, 93*(3), 765–785.

Fu, K., Cagan, J., Kotovsky, K., & Wood, K. (2013). Discovering structure in design databases through functional and surface based mapping. *Journal of Mechanical Design, 135*(3), 031006.

Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., & Wood, K. (2013). The meaning of "near" and "far": The impact of structuring design databases and the effect of distance of analogy on design output. *Journal of Mechanical Design, 135*(2), 021007.

Fujii, A., Iwayama, M., & Kando, N. (2007). Introduction to the special issue on patent processing. *Information Processing & Management, 43*(5), 1149–1153.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45.

Hedayat, A. S., Sloane, N. J. A., & Stufken, J. (1999). *Orthogonal arrays: Theory and applications*. Springer.

Jeong, C., & Kim, K. (2014). Creating patents on the new technology using analogy-based patent mining. *Expert Systems with Applications, 41*(8), 3605–3614.

Kang, I. S., Na, S. H., Kim, J., & Lee, J. H. (2007). Cluster-based patent retrieval. *Information Processing & Management, 43*(5), 1173–1182.

Kim, B., Gazzola, G., Lee, J., Kim, D., Kim, K., & Jeong, M. (2014). Inter-cluster connectivity analysis for technology opportunity discovery. *Scientometrics, 98*(3), 1811–1825.

Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications, 34*(3), 1804–1812.

Krier, M., & Zacca, F. (2002). Automatic categorisation applications at the European patent office. *World Patent Information, 24*(3), 187–196.

Kuhfeld, W. F., & Suen, C. Y. (2005). Some new orthogonal arrays. *Statistics & Probability Letters, 75*(3), 169–178.

Lamirel, J. C., Al Shehabi, S., Hoffmann, M., & François, C. (2003). Intelligent patent analysis through the use of a neural network: Experiment of multi-viewpoint analysis with the MultiSOM model. In *Proceedings of the ACL-2003 workshop on Patent corpus processing* (Vol. 20, pp. 7–23): Association for Computational Linguistics.

Lee, S., Lee, S., Seoul, H., & Park, Y. (2008). Using patent information for designing new product and technology: Keyword based technology roadmapping. *R&D Management, 38*(2), 169–188.

Lee, H. J., Lee, S., & Yoon, B. (2011). Technology clustering based on evolutionary patterns: The case of information and communications technologies. *Technological Forecasting and Social Change, 78*(6), 953–967.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6), 481–497.

Li, Y. R., Wang, L. H., & Hong, C. F. (2009). Extracting the significant-rare keywords for patent analysis. *Expert Systems with Applications, 36*(3), 5200–5204.

Liang, Y., Tan, R., & Ma, J. (2008). Patent analysis with text mining for TRIZ. In *4th IEEE international conference on management of innovation and technology, ICMIT 2008* (pp. 1147–1151). IEEE.

Loh, H. T., He, C., & Shen, L. (2006). Automatic classification of patent documents for TRIZ users. *World Patent Information, 28*(1), 6–13.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5*(2), 181–204.

Mukherjea, S., Bamba, B., & Kankar, P. (2005). Information retrieval and knowledge discovery utilizing a biomedical patent semantic web. *IEEE Transactions on Knowledge and Data Engineering, 17*(8), 1099–1110.

Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D., & Wood, K. (2014). Function based design-by-analogy: A functional vector approach to analogical search. *Journal of Mechanical Design, 136*(10), 101102.

Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications, 40*(7), 2373–2390.

Park, H., Ree, J. J., & Kim, K. (2013). Identification of promising patents for technology transfers using TRIZ evolution trends. *Expert Systems with Applications, 40*(2), 736–743.

Park, H., Yoon, J., & Kim, K. (2013). Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining. *Scientometrics, 97*(3), 883–909.

Rokaya, M., Atlam, E., Fuketa, M., Dorji, T. C., & Aoe, J. I. (2008). Ranking of field association terms using co-word analysis. *Information Processing & Management, 44*(2), 738–755.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

Salton, G., & Waldstein, R. K. (1978). Term relevance weights in on-line information retrieval. *Information Processing & Management, 14*(1), 29–35.

Schwander, P. (2000). An evaluation of patent searching resources: Comparing the professional and free on-line databases. *World Patent Information, 22*(3), 147–165.

Segers, N., & De Vries, B. (2003). The idea space system: Words as handles to a comprehensive data structure. In *Proceedings of the 10th international conference on computer aided architectural design futures digital design-research and practice*. Dordrecht: Kluwer Academic Publishers.

Seol, H., Lee, S., & Kim, C. (2011). Identifying new business areas using patent information: A DEA and text mining approach. *Expert Systems with Applications, 38*(4), 2933–2941.

Son, C., Suh, Y., Jeon, J., & Park, Y. (2012). Development of a GTM-based patent map for identifying patent vacuums. *Expert Systems with Applications, 39*(3), 2489–2500.

Souili, A., & Cavallucci, D. (2013). Toward an automatic extraction of IDM concepts from patents. In *CIRP design 2012* (pp. 115–124). London: Springer.

Souili, A., Cavallucci, D., Rousselot, F., & Zanni, C. (2011). Starting from patents to find inputs to the Problem Graph model of IDM-TRIZ. TRIZ Future.

Trappey, A. J., Trappey, C. V., Hsu, F. C., & Hsiao, D. W. (2009). A fuzzy ontological knowledge document clustering methodology. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39*(3), 806–814.

Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information, 25*(3), 211–221.

Tseng, Y. H. (2005). Text mining for patent map analysis. *Catalyst, 5424054*(5780101), 6333016.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management, 43*(5), 1216–1247.

Usui, S., Palmes, P., Nagata, K., Taniguchi, T., & Ueda, N. (2007). Keyword extraction, ranking, and organization for the neuro informatics platform. *Biosystems, 88*(3), 334–342.

Verhaegen, P. A., D'hondt, J., Vandevenne, D., Dewulf, S., & Duflou, J. R. (2011). Identifying candidates for design-by-analogy. *Computers in Industry, 62*(4), 446–459.

Wang, X., Tang, Y., & Zhang, Y. (2011). Orthogonal arrays for the estimation of global sensitivity indices based on ANOVA high-dimensional model representation. *Communications in Statistics-Simulation and Computation, 40*(9), 1324–1341.

Xie, Z., & Miyazaki, K. (2013). Evaluating the effectiveness of keyword search strategy for patent identification. *World Patent Information, 35*(1), 20–30.

Yoon, B., Lee, S., & Lee, G. (2010). Development and application of a keyword-based knowledge map for effective R&D planning. *Scientometrics, 85*(3), 803–820.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research, 15*(1), 37–50.

Yoon, B., & Park, Y. (2007). Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information. *IEEE Transactions on Engineering Management, 54*(3), 588–599.