ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Technovation 29 (2009) 481–497

technovation

www.elsevier.com/locate/technovation

An approach to discovering new technology opportunities: Keyword-based patent map approach

Sungjoo Lee[a], Byungun Yoon[b], Yongtae Park[c,*]

[a]*Department of Industrial & Information Systems Engineering, Ajou University, San 5, Woncheon-dong, Yeongtong-gu, Suwon, Kyunggi-do 443-749, Republic of Korea*

[b]*Department of Industrial & Systems Engineering, Dongguk University, Pil-dong 3 ga, Chung-gu, Seoul 100-715, Republic of Korea*

[c]*Department of Industrial Engineering, School of Engineering, Seoul National University, San 56-1, Shillim-Dong, Kwanak-Gu, Seoul 151-742, Republic of Korea*

Abstract

This paper proposes an approach for creating and utilizing keyword-based patent maps for use in new technology creation activity. The proposed approach comprises the following sub-modules. First, text mining is used to transform patent documents into structured data to identify keyword vectors. Second, principal component analysis is employed to reduce the numbers of keyword vectors to make suitable for use on a two-dimensional map. Third, patent 'vacancies', defined as blank areas in the map that are sparse in patent density but large in size, are identified. The validity of the vacancy is then tested against such criteria as technological criticality and technological trends. If a vacancy is judged as meaningful, its technological features are investigated in detail to identify the potential for new technology creation. The procedure of the proposed approach is described in detail by employing an illustrative patent database and is implemented into an expert system for new technology creation.

Keywords: Keyword-based; Text-mining; PCA; New technology creation; Patent information; Patent map; Technology vacancy

1. Introduction

Recent trends of R&D paradigms accentuate the notions of dominant designs for next-generation development, and discontinuous innovation for breakthrough products (Miller and Morris, 1999; Utterback, 1996). The creation and development of innovative ideas is recognized as an engine that gives traction to the whole innovation process in general, and new product development (NPD) and new technology creation (NTC) processes in particular. Creative ideas are the precursor of commercial success, as time and effort invested at this very early stage will pay off in the market. In fact, the impact of new products on firm success has increased markedly over time. During the 1970s, new products accounted for 20% of corporate profits, but this figure rose to 50% in the 1990s (Takeuchi and Nonaka, 1986; von Hippel, 1986). Economists have reckoned that new products produced by new technologies for new markets have resulted in between 40% and 90% of the increase in national wealth in most countries (Campbell, 1983). Today, as markets change rapidly, technologies proliferate unceasingly, and thus product life cycles become ever shorter, this factor has become more important than ever (Slater, 1993).

Despite its importance, innovation, whether for NPD or NTC, by its nature presents ill-defined and poorly-structured problems, and thus there has been little effort to systematize practice. Potential sources of innovative ideas are ubiquitous, and while internal sources have traditionally been seen as the major suppliers of new ideas, other studies argue that the majority of the ideas in many industries are derived from customers and users (von Hippel, 1986), and suppliers and even competitors may also provide insights for innovation. Consequently, a mix of internal and external resources needs to be searched (Chesbrough, 2003). The question then is how to systematically transform ideas in the form of raw data, from

*Corresponding author. Tel.: +82 2 880 8358; fax: +82 2 889 8560.
E-mail address: parkyt@cybernet.snu.ac.kr (Y. Park).

0166-4972/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.technovation.2008.10.006

whatever source, into informative knowledge for innovation. Tools such as brainstorming, group dynamics, user analysis, benchmarking, technology forecasting, or even large-scale Delphi surveys have been suggested (Urban and Hauser, 1993). But while these may direct the way we can construct organizational settings or break complex problems into simpler sub-problems, they may not offer concrete ways to facilitate creative thinking. Probably the most scientific approach to the problem of idea generation is offered by Theory of Inventive Problem-Solving (TRIZ), which uses an extensive analysis of patents to provide a tool for delicate and complex mental operations (Salamatov, 1999). While clearly both scientific and useful, TRIZ is limited in practical application by the intensive training program required.

To provide more concrete and detailed guidelines for developing innovative ideas, and especially for uncovering opportunities to create new technologies, this research suggests a keyword-based patent map approach. Patents are useful sources of knowledge about technical progress and innovative activity (Park et al., 2005) and thus have commonly been examined in R&D planning, from the macrolevel analysis of strategy to the modeling of specific emerging technologies at the microlevel (Abraham and Morita, 2001; Liu and Shyu, 1997; Wang et al., 1998; Watanabe et al., 2001). A careful analysis of the technological information in patent documents is given visualized expression as a patent map, allowing complex patent information to be understood easily and effectively (WIPO, 2003), and also highlighting various elements of knowledge about technologies, competitive positions (Abraham and Morita, 2001; Liu and Shyu, 1997), infringement risks (Daim et al., 2006), etc. Moreover, if carefully analyzed, patents can show technological details and relationships, reveal business trends, inspire novel industrial solutions and help decide investment policy (Campbell, 1983; Jung, 2003). They can also be a valuable source of information for new products or technologies. Moreover, patents are becoming increasingly important across other industry sectors, even in service sectors. Up until recently, patents, as a means to protect inventions legally, were perceived to be only for technology intensive sectors (Bader, 2008). However, as the value of firms, particularly in the knowledge-intensive business service sector, is determined by the value of their intellectual property that can be represented and protected by patents (Hanel, 2006), more firms are trying to protect their service innovations (Bader, 2008). Actually, business models and software solutions are more patentable, which have already been quite common in the US and Japan.

Given the potential utility of patent databases as sources of innovative ideas, we present a novel approach of patent analysis designed to help guide the NTC process, as well as a support system to facilitate the process of generating a keyword-based patent map to identify promising opportunities for NTC. For the purpose, we first focus on the description section of patent documents, using text-mining techniques to discover undeveloped technological fields from the patent database. In fact, there have been several attempts to apply text mining to developing patent maps (Tseng et al., 2007a, b; Yoon and Park, 2004), but previous studies have focused mainly on the development process. Studies on the interpretation algorithm have been limited, and thus interpretation of patent maps has tended to be intuitive. To overcome this limitation, our second step provides several indexes to guide patent map usage that eliminate unnecessary information so that only what is relevant to the search for next-generation technologies is sorted and revealed. This means we emphasize not only the visualization of patent map, but also its interpretation and evaluation. The suggested approach to developing patent maps relies heavily on manual work, thus restricting its operational efficiency. To solve this problem, our paper also develops a web-based software system to implement the approach proposed in this research more simply, reducing the manual work involved in information extraction, and thus allowing even those who are unfamiliar with text mining or patent analysis to benefit from the research results.

This article is organized as follows. After a brief introduction of patent analysis techniques in Section 2, the overall research process and detailed procedure of the proposed approach for NTC is described in Section 3. The approach is then illustrated in Section 4, with an exemplary patent database and is then embodied within the prototype system introduced in Section 5. Finally, some limitations of current research and suggestions for future research are discussed in Section 6.

2. Patent analysis techniques

Patents are an ample source of technical and commercial knowledge, and thus patent analysis has long been considered as a useful vehicle for R&D management. Patents possess both technical and market attributes, since they meet explicit criteria for originality, technical feasibility and commercial worth (Kuznets, 1962). However, while only few patents are developed into something of commercial value, most are technically significant in that they encourage follow-on developments in technology (Ashton and Sen, 1988). Nevertheless, patents and patenting activities are quite important for firms. It has been reported that there is a positive correlation between a firms' success and the strength of its patent portfolio (Lerner, 1994; Ernst, 2001; Shane, 2001). Ernst (1995, 2001) has found that firms that have an active patent strategy are more successful than others that remain inactive in the mechanical engineering sector. Austin (1993) has shown that patents have a positive influence on the market value of firms in the biotechnology sector. Therefore, patents are worth acquiring, investigating and analyzing. Moreover, patent databases are usually freely accessible in most counties (Daim et al., 2006) and so have advantages in terms of the availability and variety of information (Park et al., 2005). Although it is known that only patents of high value with broad technical claims and a high citation index increase the financial value of firms (Lerner, 1994; Shane, 2001), all patents have value in terms of technology development and so should be considered in analyzing their trend.

A patent document contains dozens of items for analysis, which can be grouped into two categories. The first includes structured items, which are uniform in semantics and in format across patents (such as patent number, filing date, issued date, or assignees), while the other is composed of with unstructured items, meaning they are texts of contents such as descriptions of the invention. The visualization result is called *patent graph* if an analysis of patent documents is based on the structured data and *patent map* if it is based on the unstructured data, but the general term patent maps can refer to both cases (Liu, 2003).

2.1. Structured data analysis

Although different forms of patent maps have been developed, most conventional ones use information extracted from the bibliographic fields of patent document to provide simple statistical results. For example, the Japanese Patent Office has been producing and providing more than 50 types of expressions and more than 200 maps for several technological fields (Japan Institute of Invention and Innovation (JIII), 2002), while the Korean Intellectual Property Office has plans to create maps for different technology domains over the next 5 years (Bay, 2003). Many other countries such as Italy (Camus and Brancaleon, 2003; Fattori et al., 2003) and the USA (Morris et al., 2002) develop patent maps, mostly based on the analysis of structured data from patent documents' bibliographic fields.

Likewise, patent analysis in general utilizes bibliometric data. Bibliometrics is defined as the measurement of texts and information (Norton, 2001), which helps to explore, organize and analyze large amounts of historical data in order that researchers can identify 'hidden patterns' to support their decision-making. Some common bibliometric tools employed on patent databases have been authors, affiliations, technology field, cluster and factor analysis, citations and so on (Daim et al., 2006), and one of the most frequently adopted is citation analysis (Karki, 1997; Morris et al., 2001). Patent citations are defined as the count of citations of a patent in subsequent patents, and thus citations per patent represents the relative importance of the patent.

However, conventional patent maps, while easy to understand and simple to develop, are subject to some limitations in terms of their explanatory and creative capacity, since they only use the bibliographic fields, despite the potential utility of the description section of patent documents. As a result, the scope of analysis and the richness of information are limited.

2.2. Unstructured data analysis

Recognizing the shortcomings of conventional patent maps, data mining (DM) is proposed as an alternative. DM, also known as 'knowledge discovery', is a recent development for accessing and extracting information from databases (Fayyad et al., 1996). DM applies machine-learning and statistical analysis techniques for the automatic discovery of patterns in databases, enabling the mapping of scientific and technical information by assisting in the complex process of analyzing large quantities of such information (Kim et al., 2008).

In particular, analyzing the unstructured textual data in patent documents has become possible thanks to the development of text mining (TM), a popular DM technique for handling huge amounts of unstructured textual documents (Kostoff et al., 2001). TM puts a set of labels on each document, usually by attaching them to specific words, which allows discovery operations to be performed on the labels. The text document can then be characterized according to the keywords extracted through the TM algorithm (Weiss et al., 2005). Recently, TM has attracted increasing interest and has been actively applied in patent analysis (Andal, 2006; Kim et al., 2008; Tseng et al., 2007a, b; Yoon and Park, 2004). For example, Yoon et al. (2002) applied TM to feature a patent by its keywords and suggested a new map where patents are mapped into two-dimensional space according to the similarity of their keywords. Tseng et al. (2007a, b) created a patent map for the technology domain of carbon nano-tubes based on TM. He also described a series of TM techniques including text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping. Most of those techniques are basically based on natural language processing and have been frequently applied to information retrieval (Fujii et al., 2007), summarization (Tseng et al., 2005, 2007a, b; Yoon et al., 2008), technology trend analysis (Yeap et al., 2003; Yoon and Park, 2004; Yoon et al., 2002), and automated classification applications (Schellner, 2002; Krier and Zaccà, 2002; Larkey, 1999). Especially, automatic methods for patent classification are expected to relieve the work by human experts, which is expensive and subjective (Fujii et al., 2007). Consequently, a number of organizations including the The United States Patent and Trademark Office (USPTO) (Larkey, 1999), the EPO (Krier and Zaccà, 2002), the JPO (Delegation of Japan, 2000), INPI (Leclercq, 1999; Lyon, 1999), and Xerox (Hull et al., 2001) are tying to design and implement systems of automated classification of patent documents.

The most important advantage of applying TM in patent analysis is that it can handle large volumes of patent documents and extract some meaningful implications from textual data. Patent documents contain significant research outputs but are so lengthy that a lot of human efforts are required to analyze the contents. TM will assist patent engineers or decision-makers in patent analysis (Smith, 2002; Tseng et al., 2005, 2007a, b). In addition, TM might be better than human in some areas. Tseng et al. (2005) provide that the machine-derived important categorization features might be as good as or even better than those derived manually. Though software based on TM can be difficult to make proper distinctions in others and thus may not be acceptable for some areas where accuracy is emphasized (Smith, 2002), it is very useful at least for pre-classification and supportive information (Krier and Zaccà, 2002).

However, patent documents pose unique challenges to TM application. Firstly, regarding the TM algorithm, most exiting TM software cannot consider synonyms and/or co-occurrences of keywords (Uchida et al., 2004). In addition, TM algorithm does not include compound words due to the difficulties of determining them, which is critical to improve the algorithm. Also, relatively large numbers of keywords are required to ensure a subtle distinction between documents using TM (Smith, 2002). Secondly, relating to the feature of patent documents, the claims, which precisely specify the boundary of the invention and thus are valuable for TM, are generally written in arcane legalese and thus difficult to extract technical meanings. Moreover, when the whole body of patent document is used for TM, it is not easy to distinguish texts describing ''prior art'' from texts describing ''the invention'', which is essential to address the technical characteristics of the invention (Smith, 2002). Despite those limitations, advances in computer technologies are expected to increase the efficiency and accuracy of TM, strengthening the advantages.

In this manner, therefore, various studies have been conducted to analyze unstructured patent information, but little was concentrated on applying related TM techniques in the context of NTC or NPD. Besides, 'interpretation' during patent processing, is relatively at the early stage of research compared to 'information retrieval' or 'classification', though it is more difficult (Fujii et al., 2007). Therefore, this research adopts TM to analyze the technological contents of patent documents, and develops a keyword-based patent map where patents that are similar in terms of technological contents are located close to each other, as does the existing research. However, the suggested map is then used as a basis for NTC and its interpretation is emphasized to maximize its use.

2.3. Patent analysis systems

Patent analysis to extract useful knowledge from databases is usually time consuming, which requires thousands of patents to be examined and information classified. Consequently, manual generation of patent map is very costly and has a limited supply (Uchida et al., 2004). Therefore, the automation of patent map generation and analysis is in great demand. To reflect the need, several software systems have been developed, out of which some provide only a function to generate ''patent graph'', while the others adopt TM algorithm to help generate ''patent map'' as well. To name a few, Thomson (http://scientific.thomson.com/products/aureka) has introduced its advanced Intellectual Property (IP) management and analysis system, called Aureka, which is characterized by such features as messaging, annotations, proprietary fields, alerts, themes cape mapping and hosted/local model. Neopatents (http://www.neopatents.com) has developed two software systems specifically designed for patent search and analysis—one is Spore® Search that optimizes the search process based on statistical permutations of the keywords and the other is PatentMatrix® that provides graphical representations of analysis results. Metrics Group (http://www.metricsgroup.com) provides three different systems useful for patent analysis: (1) CIA™ Database that is a web-enabled system to discover trends and patterns in US patent databases; (2) text mining software called VantagePoint™; and (3) VxInsight™ developed as a visualization tool to explore patent data. Other systems have been suggested to address the specific function of searching, which includes IPMap (http://www.dolcera.com), Invengine™ (http://www.invengine.net), M-Cam Door (http://www.m-cam.com). They can function as an IP search engine, layering onto patent and technology databases and enabling accurate and precise searching with results in an effective and easy to review format. Along with those efforts in practical areas, academic areas continue to struggle with problems in the automatic generation of patent maps, increasing the effectiveness and efficiency of algorithms applied. For example, Fattori et al. (2003) adopted TM technology to overcome the limits of existing patent classifications approaches. They suggested a bottom-up clustering approach instead of a top-down classification for patent categorization and developed the system called PackMOLE™, though insisting those two approaches should be used in synergy. Later, not only TM but also DM was investigated to gain information for strategic business decisions. Fischer and Lalyre (2006) developed STN®AnaVist™ using host-based analysis and visualization, where many interactive analysis and visualization options are available and the navigation between the individual data elements is easy. Tables 1 and 2 summarize the representative systems of patent analysis with their features.

As is shown in the tables, firstly in the perspective of *process*, most well-known systems have focused on patent search, analysis, and visualization, though the utility of some systems are not limited to patent documents. They are quite useful for an effective analysis of current states of data and have proved their value in a real business scenario. However, the systems concerning knowledge interpretation issues are rare, which can help to analyze patent mapping results and is more difficult but essential to ensure the benefits of patent mapping. Since the techniques for the patent analysis and mapping have developed dramatically and thus the further research should integrate some kinds of facilities for manipulating patent mapping results and other descriptive indexes to interpret the results, to speed up the whole process and to guarantee that even people who are not familiar with patent map can get the benefit from the systems. Secondly in the perspective of *context*, most existing systems are aiming to support the strategic decision-makings at the management level, while few are developed to support the practical affairs of engineers. Main applications of the exiting systems have been business, technology, and especially IP management to maximize the return of investment in patents. It is true that IP management is an important issue in patent map. However, another promising research area for patent analysis can be the contents analysis to support the work by engineers, since patent documents contain detailed information on inventions and technologies and thus can be useful knowledge sources for NTC. To tackle missing functions that existing systems have not covered, the system suggested in this research focuses on the interpretation stage of patent analysis and is designed to support decision-makings at the engineering level rather than at the strategic level, which can enrich the functions of existing systems playing the complementary role and thus be adopted as their sub-modules.

Table 1
The existing systems for patent analysis.

System	Target	Source
Aureka	Documents	Thomson (http://scientific.thomson.com/products/aureka)
IPMap	Documents	Dolcera (http://www.dolcera.com)
PatentMatrix®	Patents	Neopatents (http://www.neopatents.com)
Spore® Search	Patents	Neopatents (http://www.neopatents.com)
Invengine™	Patents	Invengine (http://www.invengine.net)
M-Cam Door	Documents	M-Cam (http://www.m-cam.com)
CIA™ Database	Patents	Metrics Group (http://www.metricsgroup.com)
VantagePoint™	Documents	Metrics Group (http://www.metricsgroup.com)
VxInsight™	Patents	Metrics Group (http://www.metricsgroup.com)
ATMS/Analyzer	Patents	Fujitsu (http://glovia.fujitsu.com)
PackMOLE™	Patents	Fattori et al. (2003)
STN®AnaVist™	Documents	Fischer and Lalyre (2006)

2.4. Keyword-based patent map

Keyword-based patent maps place patents into a two-dimensional space according to similarity of keywords in each patent. To develop the patent map, this study applies two DM techniques—TM and principal component analysis (PCA). First, since patent documents per se are expressed in unstructured text, systematic methods are required to perform knowledge discovery from content in this format, and TM is adopted for this data structuring. The other major tool is a two-dimensional visualization algorithm to create patent map. Patent maps in general are defined as a visual form—a chart, table or graph—that analyzes or arranges patent data to make it more informative and constructive. Since TM merely generates document with keywords, it is of little help in interpreting technological significance and drawing strategic implication. Patent maps assist users in grasping diverse features of individual patents and identifying complex relationship among patents. However, since patent maps can consist of so many dimensions of variables as to make them difficult to comprehend, one of the critical tasks is to reduce the number of dimensions of keywords to acceptable levels, for instance for a two-dimensional map.

For the mapping, several mapping methods, from the basic Vector Space Model (VSM) (Salton and Bucklye, 1998) to more advance ones such as the Self-Organizing Map (SOM) neural network technique (Morris et al., 2001; Yoon et al., 2002), force-directed placement (FDP)

Table 2
Features of the existing systems for patent analysis.

System	Main functions				Main techniques	Main applications
	S	A	M	I		
Aureka	●	●	●		Web-based, simple statistics, data mining, text mining, citation analysis	IP strategy
IPMap	●				Web-based, simple statistics	IP strategy
PatentMatrix®		●	●		claim analysis	IP strategy
Spore® Search	●		●		Semantic-based statistical search	IP search and mapping
Invengine™	●				Web-based, semantic analysis	IP search
M-Cam Door	●				Linguistic genomic algorithm	IP risk management
CIA™ Database	●				Web-based, simple statistics, data mining, citation analysis	R&D and business strategy
VantagePoint™		●			Semantic analysis	Text mining
VxInsight™			●		Data mining, 3-D data visualization	IP strategy, R&D and business strategy
ATMS/Analyzer		●			Simple statistics, data mining	R&D and business strategies
PackMOLE™		●			Text mining, data mining	Patent clustering
STN®AnaVist™		●	●		Host-based, simple statistics, text mining, data mining	R&D and business strategies

Note: S: search (collection), A: analysis, M: mapping (visualization), I: interpretation

(Davidson et al., 1998), Latent Semantic Indexing (LSI), have been proposed (Deerwester et al., 1990), but their utility is limited since multi-dimensional information is decomposed into two "unclear" dimensions. Correspondence analysis (CA) might be applied, which is one type of PCA, yielding joint graphical displays and using a contingency table for categorical variables (Greenacre, 1984). Its distinctive strength over other methods is that it produces two dual displays, whose row and column geometries have similar interpretations, while other multivariate approaches to mapping do not have this duality (Theodorou et al., 2007). However, this research treats the keyword frequency that will be the basis of mapping as continuous values, not as categorical values, and also focuses only on the relationships between documents and is not interested in the relationships between keywords. Another is a co-occurrence based model (Schűtze and Pedersen, 1994, 1995; Uchida et al., 2004), but again, it does not deal with the word frequencies that may cause information loss. As a result, we concluded that the most suitable method for this research will be PCA, whose general objective is data reduction by converting different variables into a few linear combinations (Johnson and Wichern, 1998).

PCA has four main advantages over other methods. First, the meaning of axes in the two-dimensional map can be interpreted easily and clearly since it reveals the relationship between each principal component and its respective variables. Second, it provides richer information, such as expressing the absolute location of each patent as a numerical value. Third, unlike the SOM technique, all patents can be represented on a single map. Finally, we can get several PCs from the PCA results, out of which various combinations of two PCs generate several patent maps. Since the purpose of this research is not to map patents on a two-dimensional space accurately but to explore new opportunities through the mapping, PCA that enables to develop several patent maps in various perspectives is judged to be more suitable techniques.

The basic idea of our proposed approach for discovering NTC opportunities using keyword-based patent map is as follows. If we can excavate the latent characteristics of patents and locate individual patents on the map using these characteristics, we may be able to identify some 'vacancies' in the map. In this context, a vacancy can be defined as a blank zone surrounded by many existing patents. It therefore represents an as yet unexplored area, but one which may have development potential for the future, given the active development of adjacent areas. Such technologically undeveloped zones can have great NTC potential, as successful occupation of such vacant areas may offer significant first mover advantages. Promising opportunities for NTC may then be derived by intensively examining surrounding patents.

In spite of the problems inherent in applying TM in patent analysis, we judge that it is not so critical in this research. Since the focus of this research is the technical contents rather than the right of the invention, we use the abstract or the whole body instead of the claims. It enables us to eliminate the problems caused by failing to extract enough keywords to distinguish documents or having difficulties in analyzing a legal term of claims. Nevertheless, additional work by experts was designed to solve the problem. For example, extracted keywords are encouraged to be reviewed by experts with domain knowledge, considering synonymies and compound words before determining the final set of keywords to be used for the further analysis.

3. Finding new technology opportunities

3.1. Overall research framework

The overall process of our proposed approach consists of the following three modules—*development of patent map*, *identification of patent vacancy*, and *test of vacancy validity*. These major modules are each composed of several sub-modules that carry out more detailed functions. Fig. 1 shows the major modules and detailed sub-modules of the overall process.

In the first phase, patents are located on a two-dimensional map according to their technological contents. Then at the second phase, we identify patent vacancies from the map, which aids the identification of any technologically undeveloped 'blank zones'. The final phase involves screening to identify and investigate meaningful patent vacancy areas: as not all will be worth investigating for potential new technology development, they can be classified as either 'fruitful' or 'barren' areas. For this purpose, two further analysis types—*criticality analysis* and *trend analysis*—are conducted. In the criticality

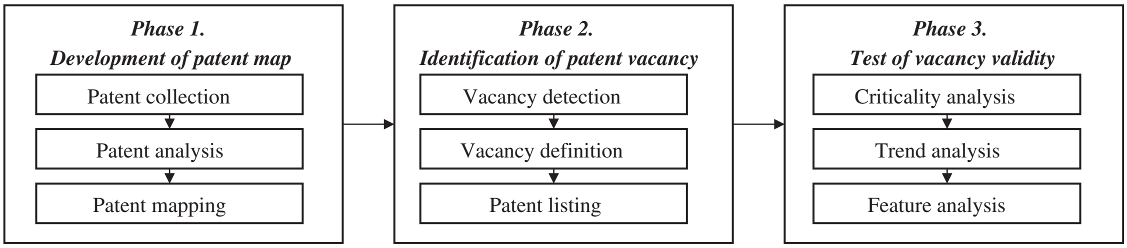


Fig. 1. Overall process of NTC approach.

analysis, patents in areas around the vacancy (we call them *adjacent patents*) are examined on the basis of citation, claims and density, which may tell us whether the vacancy is technologically valuable or not. Trend analysis measures the percentage of keywords and classes appearing in adjacent patents that are associated with the latest technologies, which may tell us whether the vacancy is associated with emerging or declining technologies. Where the analysis results suggest the vacancy has valuable potential for NTC development, we recommend deeper investigation of patents in the adjacent areas, to see if analyzing, recombining and integrating the patents can identify NTC opportunities. More detailed explanation of the process is provided below.

3.2. *Detailed procedures*

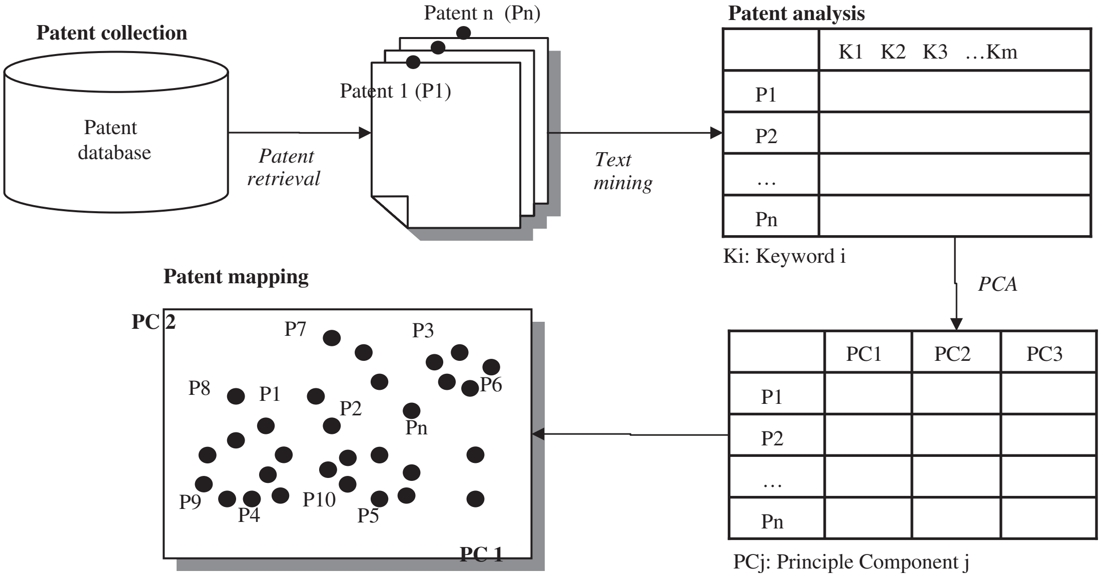
3.2.1. *Development of patent map*

Fist, patent map is developed through three successive steps—patent collection, patent analysis and patent mapping (See Fig. 2).

Patent collection: The first task is to collect patent documents in the specific technology field to be analyzed. Recent advances and diffusion of internet-based abstract services allow easy access to patent databases in electronic form. Patent documents collected at this stage are unstructured data, in that they are merely expressed in text format.

Patent analysis: The next step is to transform unstructured the text document into structured data. A typical patent analysis scenario includes task identification, searching, segmentation, abstracting, clustering, visualization, and interpretation (Tseng et al., 2007a, b). Out of them, as described above, TM serves as a tool for analyzing relationships among patents to be used a basis for the clustering, visualization, and interpretation. Specifically, each patent will include keywords that can be used to represent its technological characteristics. Here, how to extract keywords to be used for the analysis relies firstly on TM software and secondly experts' judgment. TM software can yield an importance of each keyword in the whole documents. Or the keyword frequency in the whole documents can be used as a proxy measure of the importance. Then, keywords only with high importance will be selected as the first candidates and go through experts' screening. The remaining keywords after the screening will be a final subject of analysis. Then, for each patent, the frequency of the keywords' use in the patent documentation is assigned to a corresponding vector field, and thus each patent document can be distinguished by a keyword vector.

Patent mapping: Once the keyword vector is completed, documents are mapped to a rectangular planar surface in order to generate the patent map. In applying PCA to creating the map, an important task is to determine the number of components. While there is no definitive method to determine the optimal number of components, the components that can well-explain the variance of overall variables are used in a business practice and usually three or four may be sufficient to explain the majority of total sample variance. This explanation power can be measured by the eigenvalue of each component. In general, the components whose eigenvalue is over one are regarded as principle components (PCs), while close to zero should be excluded from them (Johnson and Wichern, 1998). Therefore, information loss might be a problem. To deal with it, using more than two PCs, for example, developing a three-dimensional map based on three PCs or examining several patent maps based on various combinations of two PCs, is encouraged.



	K1 K2 K3 …Km
P1	
P2	
…	
Pn	

	PC1	PC2	PC3
P1			
P2			
…			
Pn			

Fig. 2. Process of patent map development.

3.2.2. Identification of patent vacancy

After patent map is generated, the second module focuses on identifying patent vacancy areas through three steps—vacancy detection, vacancy definition, and patent listing (see Fig. 3).

Vacancy detection: As a preliminary step, the patent map is examined to detect sparse areas, defined as relatively large areas where patent density is extremely low. Although quantitative criteria can be applied to detect possible areas for further investigation, qualitative and intuitive judgments may be more flexible in terms of identifying areas with real potential.

Vacancy definition: In general, sparse areas on patent maps will be surrounded by many patents: however, it is unrealistic to consider all the surrounded patents, and a subset of related patents must be used to determine the 'boundary' of the patent vacancy. The criteria for the subset and the determination of which adjacent patents to examine will be subjective to the investigating company, and the process is of necessity conducted manually. For example, if a company is conducting an exploratory research to discover the NTC possibilities by and large, more information may be provided by including more patents in adjacent patent group. Whereas, if a company is interested in minor innovation to avoid infringement risk, restricting analysis scope to the patents directly bounding on the vacancies will give more practical solution. Analysis range, of course, will be affected by the capability of the company to retrieve the patents. Once a sparse area on the matrix is detected and the surrounded patents selected, a patent vacancy is generated. It is common that a number of vacancies can be drawn from a given patent map.

Patent listing: After developing the patent map and identifying patent vacancies, the next step would be to investigate the validity of each vacancy. The validation step is indispensable, since some vacancies may appear to be fertile but turn out to be sterile in terms of the potential value of surrounded patents. The validation criterion is based on the idea that the importance of a vacancy is determined by some primary characteristics of the surrounding patents. Thus, for each vacancy, the task required for validation is to organize the list of patents that are both adjacent to each other and located on the vacancy boundary.

3.2.3. Test of vacancy validity

The final module is designed to test the validity of each vacancy identified in the previous module through three steps—criticality analysis, trend analysis, and feature analysis (see Fig. 4).

The first task for the validation is to collect primary information for each of patents that are used to define the vacancies. The scope of information is wide and diverse, ranging from basic information (patent number, title and assignee, etc.) to more sophisticated information such as abstract, description and claim. Some of this information is used to calculate the development potentials. Among others, the following three indexes for criticality analysis and four indexes for trend analysis are operationally defined. Then, based on those indexes, the vacancies to be explored are finally determined and their technological features are analyzed to find NTC opportunities.

Criticality analysis: Criticality analysis aims at evaluating criticality of each vacancy for future NTC activities. In this analysis, the degree of criticality of adjacent patents is measured by the average frequency of citation, the average number of claims, and the density of the vacancy. Firstly, *citation* measures how often the patent is cited in other patents. Since patent citations have significant relationship with patent value (Engelsman and van Raan, 1994; Karki, 1997; Lanjouw and Schankerman, 1999), the average number of citations of adjacent patents is regarded as the indicator for the importance of the vacancy. This is measured in the Value of Technology (VoT) index. Secondly, *claim* is defined as the average number of claim items per patent. The claims specify in detail the building blocks of the patented invention and the number may be indicative of the scope or width of the patent (Ernst, 2003; Lanjouw and Schankerman, 1999; Park et al., 2005). This leads to the Scope of Technology (SoT) index, which measures the average number of claim items of adjacent patents. Finally, *density* is calculated as the number of all adjacent patents divided by the size of the vacancy area in the Competition of Technology (CoT) index. A high value, which indicates keen competition between adjacent patents (Kohonen, 1995; Yoon, 2005), can hint that the vacancy area has a rich NTC.

Trend analysis: This analysis evaluates adjacent patents in terms of how they relate to the most up-to-date

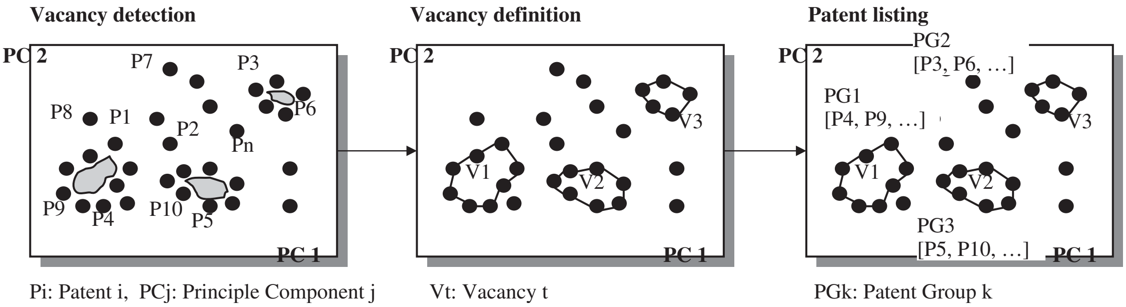
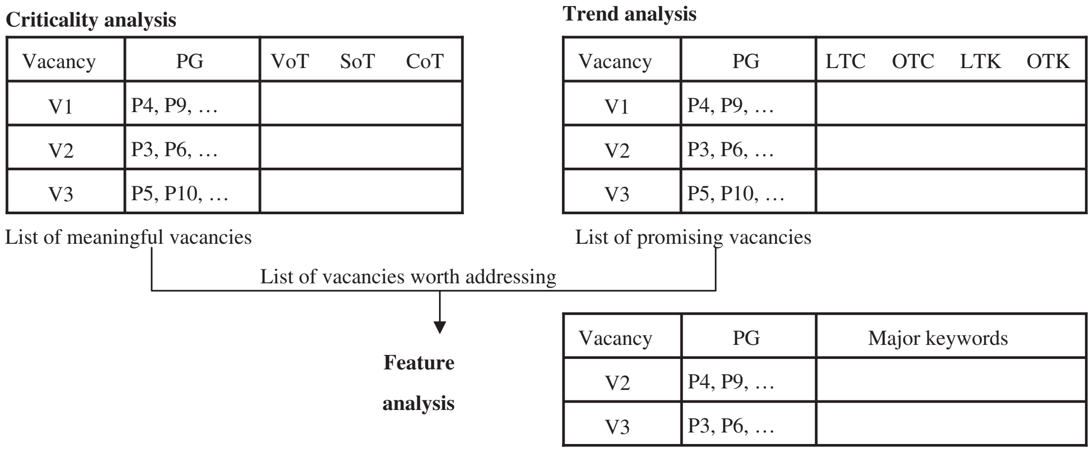


Fig. 3. Process of patent vacancy identification.


Criticality analysis
Trend analysis
List of meaningful vacancies
List of promising vacancies
List of vacancies worth addressing
Feature analysis

Vacancy	PG	VoT SoT CoT
V1	P4, P9, …	
V2	P3, P6, …	
V3	P5, P10, …	

Vacancy	PG	LTC OTC LTK OTK
V1	P4, P9, …	
V2	P3, P6, …	
V3	P5, P10, …	

Vacancy	PG	Major keywords
V2	P4, P9, …	
V3	P3, P6, …	

Fig. 4. Process of vacancy validity test.

Table 3
Indexes to test vacancy importance.

Purpose	Index	Information source	Definition
Significance analysis	VoT (the degree of technological diffusion)	Citation	The average number of citations of adjacent patents
	SoT (the scope of legal protection)	Claim	The average number of claim items of adjacent patents
	CoT (the degree of technological competition)	Number	The number of all adjacent patents divided by vacancy size
Trend analysis	LTC (the latest technological contents)	Abstract	The percentage of patents having emerging technological keywords out of all adjacent patents
	OTC (the outdated technological knowledge flow)	Abstract	The percentage of patents having declining technological keywords out of all adjacent patents
	LTK (the latest technological knowledge flow)	Class, citation	The percentage of patents citing emerging classes out of all adjacent patents
	OTK (the outdated technological knowledge flow)	Class, citation	The percentage of patents citing declining classes out of all adjacent patents

technologies and technological knowledge flow. The first—trend analysis of technological contents—is based on the extent to which the keywords can be categorized as either of two types: emerging or declining keywords. A time-series analysis of keyword frequency in technological documents can show general features of technology trend (Lee et al., 2008; Yoon et al., 2008) and thus by comparing the analysis results with technological contents in adjacent patents of a vacancy, we are expected to measure its importance in terms of technology trend. In specific, the higher the percentage of adjacent patents with emerging keywords, and the lower the percentage with declining keywords, the more the vacancy reflects the latest technological trends. On the other hand, the second—trend analysis of technological knowledge flow—is based on the class information. When patents are issued, classification codes are used to assign them to similar technology groups. Here, emerging and declining classes are defined according to the rate of increase in patent applications in these classes. If patents adjacent to a specific vacuum are citing more patents coded to emerging classes and fewer to declining classes, the vacancy can be judged as reflecting the most recent flows of technological knowledge on the assumption that knowledge flows in patents represent technological flows (Scherer, 1981). For the purpose, four indexes—Latest Technological Contents (LTC), Outdated Technological Contents (OTC), Latest Technological Knowledge-flow (LTK), and Outdated Technological Knowledge-flow (OTK)—are designed to measure the extent to which the vacancy reflects the most recent technological trends. While LTC and OTC measure the percentage of adjacent patents which have emerging or declining technological keywords, LTK and OTK measure the percentage of adjacent patents citing emerging or declining classes. The operational definition of indexes to test vacancy importance is described in Table 3.

Feature analysis: The final task is to perform a detail analysis of the features of those vacancies that have been identified as meaningful and promising. For vacancies thus selected, all keywords related to adjacent patents are collected to infer their technological features and seek for significant opportunities for new technology development.

4. An example of approach implementation: the case of PDA technology

In the current research, patents related to Personal Digital Assistant (PDA) technologies are employed for illustration. The United States Patent and Trademark Office (USPTO) database serves as the source for collecting patent documents. In all, 141 PDA-related patents are gathered covering the period 1996–2003. The set ranges from US Patent No. 5 497 339–6 516 251, but since real patent numbers are too cumbersome for database analysis and map display, they have been given serial numbers (1–141) according to their application date. The reason why PDA-related patents were selected is two-fold. First, PDA technology is suitable for monitoring trends due to its rapid technical advancement, which facilitates the diffusion of mobile services. Second, the PDA database is of a convenient size for mining latent information and mapping in two-dimensional space.

4.1. Patent map development

4.1.1. Development of keyword vector program

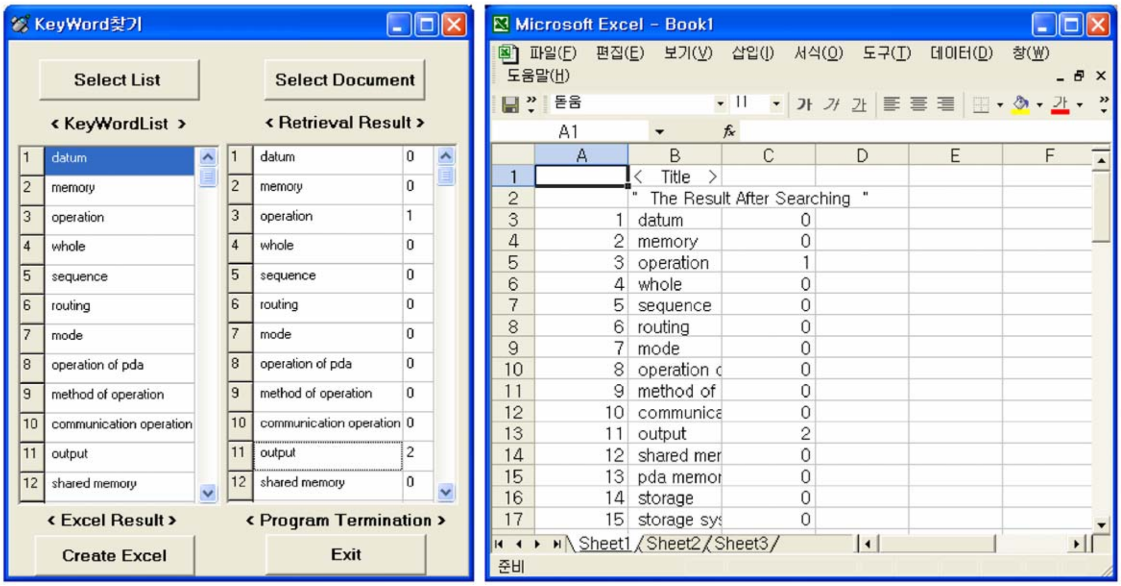
The output of TM for the 141 PDA-related patent documents is a keyword list of all patents in an Excel file. Considerable time and effort has previously been required to change the unstructured text file into a structured excel file as few tools have been developed for such transformation purposes. To remedy this situation, we have developed a keyword vector construction tool using a programming language Delphi. First, users select the keyword list which should be prepared in Excel format and the documents in Text format. Then, when the 'Create Excel' command button is clicked, the frequency table is built in Excel format and counts are displayed in the 〈Retrieval Result〉 list box in the order of keyword list file (see Fig. 5). The left figure indicates that the keyword "operation" appears twice in the first document.

4.1.2. Selection of keywords

After executing TM, the keywords are arranged in the form of a hierarchical structure. In this case, the 141 patent documents yielded 39 keywords, which can be arranged on a six-level hierarchy. Basically, two different approaches are possible in selecting keywords. One method is an 'all-keyword' approach that utilizes all the keywords in constructing the keyword vector. The other is the 'major-keyword' approach that uses only a selective set of the most significant keywords. The former may reduce information loss, but the latter may improve the explanatory power of the process by employing fewer PCs. Based on a number of experimental tests, we found that the difference between the two approaches is negligible, implying that the mapping structure is quite robust to the number of keywords (see Appendix A), but the former can keep more information than the latter and its explanation power for the same number of PCs is superior to that of the latter as well. Thus, the all-keyword approach is adopted in this research.

4.1.3. Construction of keyword vector

With the aid of a keyword vector construction program, 141 keyword vectors for individual patent documents are constructed, as illustrated in Fig. 6. The keyword vector fields register the frequency of keyword occurrence: thus a frequency of 3 for the 'datum' field in Patent 1 means that the keyword 'datum' occurs three times in the Patent 1 documentation.



1	datum
2	memory
3	operation
4	whole
5	sequence
6	routing
7	mode
8	operation of pda
9	method of operation
10	communication operation
11	output
12	shared memory

1	datum	0
2	memory	0
3	operation	1
4	whole	0
5	sequence	0
6	routing	0
7	mode	0
8	operation of pda	0
9	method of operation	0
10	communication operation	0
11	output	2
12	shared memory	0

	A	B	C	D	E	F
1		< Title >				
2		" The Result After Searching "				
3	1	datum	0			
4	2	memory	0			
5	3	operation	1			
6	4	whole	0			
7	5	sequence	0			
8	6	routing	0			
9	7	mode	0			
10	8	operation c	0			
11	9	method of	0			
12	10	communica	0			
13	11	output	2			
14	12	shared mer	0			
15	13	pda memor	0			
16	14	storage	0			
17	15	storage sys	0			

Fig. 5. An illustration of keyword vector construction program.

Keyword	:		datum,	memory,	operation,	sequence,	•••	electronic datum	
Patent 1	:	(	3 ,	4 ,	0 ,	0 ,	•••	0	)
Patent 2	:	(	1 ,	0 ,	0 ,	0 ,	•••	0	)
•							⋮		
Patent 141	:	(	2 ,	5 ,	3 ,	31 ,	•••	0	)

Fig. 6. Keyword frequency vector.

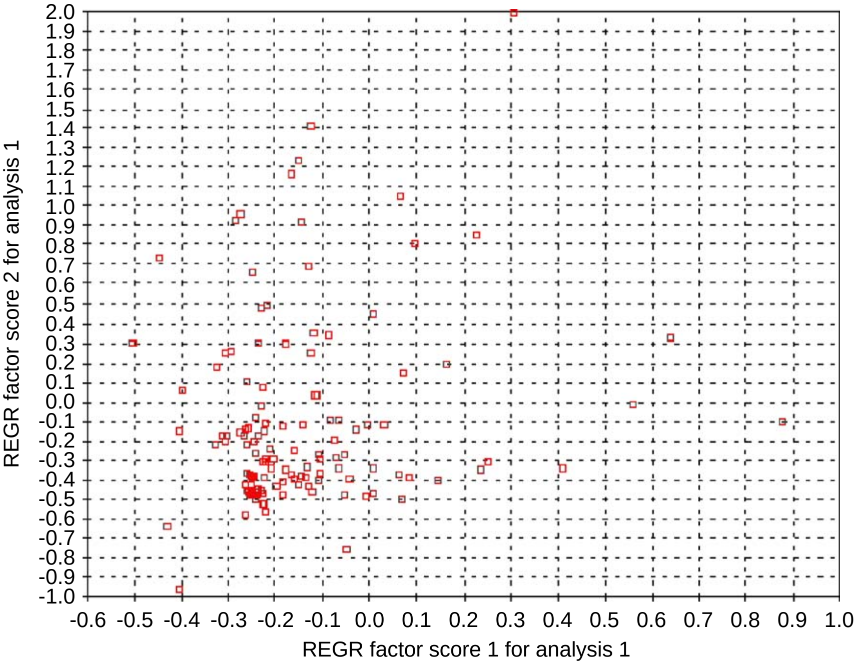


Fig. 7. An example of two-dimensional patent map.

4.2. Patent vacancy identification

After data preprocessing, various patent maps may be developed based on the PCA results of the keyword vector, which identified several PCs. The implication of each PC can be inferred from the meaning that all keywords having high factor loadings on the PC have in common. In this research, however, too many keywords were related to one PC and thus it was not easy to extract any common meaning. Moreover the purpose of mapping is to find vacancies, not to interpret the meanings on the map and thus the description of PCs was not analyzed. If there are not so many keywords used for the analysis and the interpretation of map is required, it would be possible to give some meanings to the PCs. Fig. 7 demonstrates an example of a patent map comprising two PCs, PC1 and PC2.

In Fig. 8, the ellipses in the patent map on the left-hand side indicate patent vacancies that can be identified visually. The polygons in the patent map on the right-hand side exhibit vacancies detected by connecting surrounding patents. In all, six vacancies, labeled 1–6, are generated.

As pointed out before, adjacent patents are listed for each vacancy before subsequent analysis is conducted to test the importance of the vacancy. Table 4 shows the list of adjacent patents for the six vacancies. The number of adjacent patents range from a minimum of 7 to a maximum of 15.

4.3. Vacancy validity test

4.3.1. Validity test results

Each vacancy is then subject to criticality analysis and trend analysis. The criticality analysis results are summarized in Table 5.

Note that the validity and potential utility of a vacancy is determined by the values of the above indexes. If a vacancy has higher values for all indexes, it is evaluated as a meaningful vacancy. To illustrate, vacancy 3 has the highest values in all three indexes and thus definitely warrants intensive analysis. In case of vacancies 1, 4 and 5,

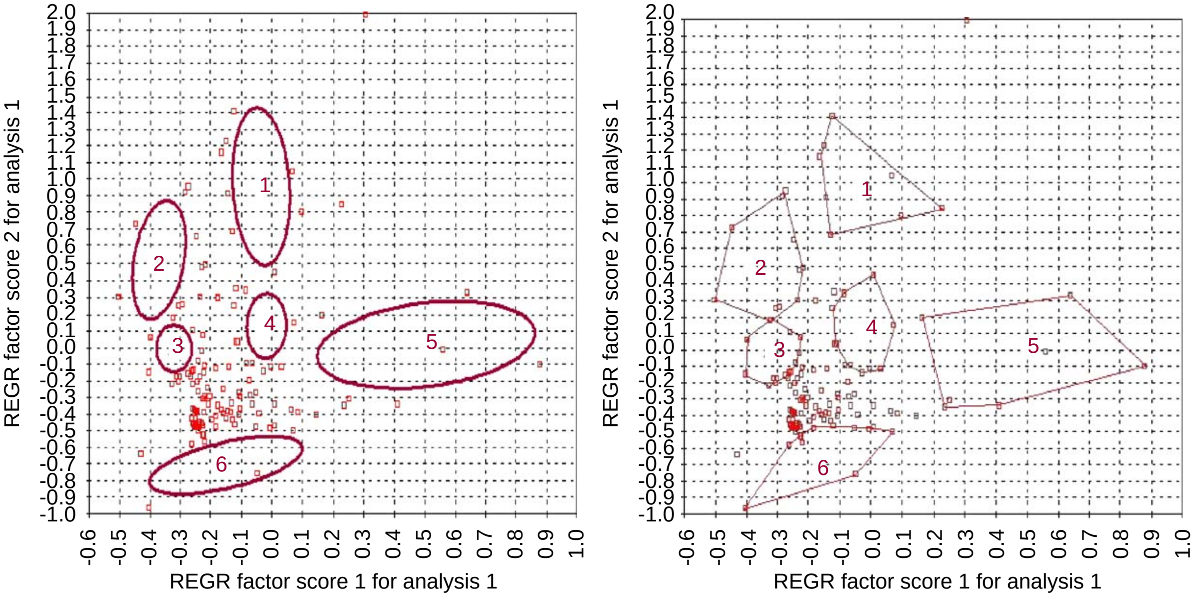


Fig. 8. Illustrations of patent vacancies.

Table 4
List of adjacent patents for patent vacancies.

Vacancy 1	Vacancy 2	Vacancy 3	Vacancy 4	Vacancy 5	Vacancy 6
Patent list					
5 675 524	5 778 256	4 083 797	5 600 841	5 704 029	5 515 305
5 687 254	5 938 721	4 612 584	5 797 089	5 742 905	5 802 275
6 016 476	5 983 073	5 526 481	5 952 994	6 034 621	5 974 334
6 049 453	6 018 724	5 606 594	6 032 255	6 169 498	6 101 562
6 137 481	6 266 612	5 699 244	6 173 310	6 189 056	6 163 274
6 308 201	6 317 718	5 819 227	6 292 186	6 343 148	6 195 589
6 457 062	6 411 899	6 168 331	6 301 551	6 403 312	6 230 303
6 459 969	6 421 235	6 198 941	6 356 956		6 233 464
	6 424 369	6 244 873	6 360 172		6 247 947
	6 467 088	6 270 271	6 366 898		6 262 684
	6 475 146	6 305 603	6 421 232		6 334 575
		6 360 172			6 363 271
		6 366 450			6 393 463
		6 424 369			6 516 251
		6 512 515			
Total number					
8	11	15	11	7	14

Table 5
Criticality analysis results.

	Vacancy 1	Vacancy 2	Vacancy 3	Vacancy 4	Vacancy 5	Vacancy 6
VoT	3.50	1.00	13.33	0.67	10.00	0.67
SoT	12.00	3.00	26.88	26.50	17.00	2.00
CoT	0.57	0.85	3.00	1.47	0.23	1.27

some of indexes show relatively high values, and thus have some potential and deserve in-depth investigation. However, vacancies 2 and 6 have relatively low values for all indexes and seem to display little potential. Fig. 9 indicates the relative importance of respective vacancies in the map.

Trend analysis is then conducted. With respect to the trend of technological contents, emerging keywords include ‘mode’, ‘store’, rom’, ‘ICMCIA’ and so on, while declining keywords include ‘routing’, ‘workstation’, ‘PC software’, etc., while for the technological knowledge flow trend, emerging classes include 345, 382, 435, 702, 707, 709, and 712, while declining classes include 235, 365, 378, 432, and 456. This trend is based on the 141 PDA-related patent documents. Emerging (declining) keywords are the keywords whose appearance frequency in the documents is increasing (decreasing) and the classes where the number of patents applied to the class is increasing (decreasing) are defined as emerging (declining) classes. Table 6 shows trend analysis results giving the percentage of adjacent patents which have emerging/declining keywords, and which cite patents from emerging/declining classes for each vacancy.

Table 7 shows the final results of the vacancy validity tests, putting together the criticality analysis and trend analysis results.

As is shown in Table 5, those vacancies judged as critical by the criticality analysis are not absolutely in keeping with those identified by the trend analysis. This is because criticality analysis regards patent vacancies with adjacent patents that are frequently cited and have many claim items as a critical, while trend analysis regards patent vacancies that reflect the latest technological trends as the valuable ones. Since emerging areas, of course, have not yet been exposed to development competition, it is likely that patents in these areas show relatively lower values for frequency of citation and number of claims. Thus vacancy 3 can be considered a *critical* but *declining* vacancy. It means that its adjacent patents tend to be old ones which

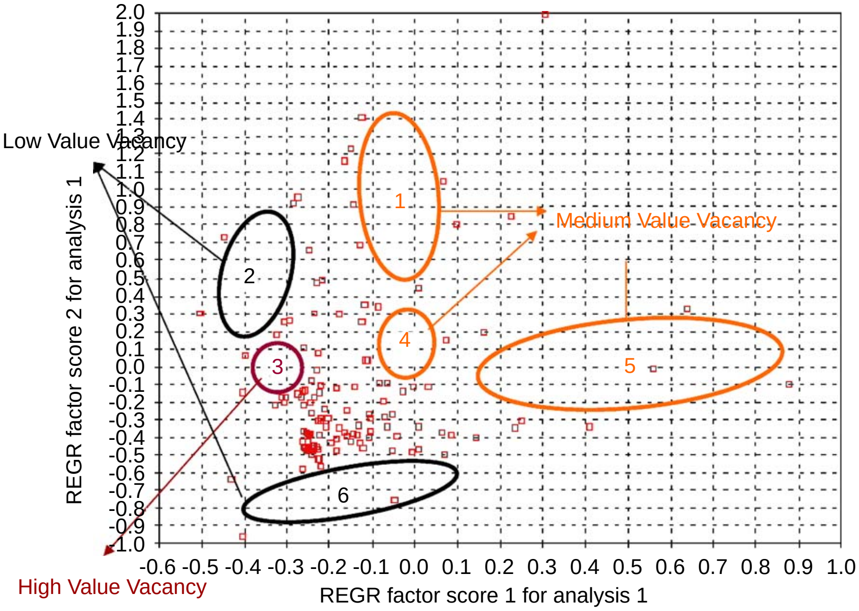


Fig. 9. Visual identification of significant vacancies.

Table 6
Trend analysis results.

	Vacancy 1	Vacancy 2	Vacancy 3	Vacancy 4	Vacancy 5	Vacancy 6
Technological contents						
Emerging keywords	0.813	0.932	0.733	0.886	0.893	0.768
Declining keywords	0.344	0.523	0.323	0.364	0.250	0.268
Technological knowledge flow						
Emerging classes	0.232	0.234	0.162	0.312	0.245	0.184
Declining classes	0.100	0.110	0.173	0.127	0.135	0.167
Evaluation results	Emerging	Neutral	Declining	Emerging	Emerging	Declining

Table 7
Final results of vacancy validity test.

	Vacancy 1	Vacancy 2	Vacancy 3	Vacancy 4	Vacancy 5	Vacancy 6
Criticality analysis	Meaningful	Meaningless	Critical	Meaningful	Meaningful	Meaningless
Trend analysis	Emerging	Neutral	Declining	Emerging	Emerging	Declining

involve what has have been regarded as a major technology hitherto. If a firm wishes to develop a core competence, relying on comparatively established but newly-structured technologies, and needs to avoid patent infringement, finding NTC opportunities though vacancy 3 would be beneficial. On the other hand, vacancies 1, 4, and 5 are characterized as *meaningful* and *emerging* vacancies. Most of their adjacent patents correspond to the latest technological trends, and have been frequently cited, with large numbers of claim items, and show a fairly high density. If a firms wishing to identify emerging technologies which are not core yet but appear likely to be significant in the future, and is willing to take the risk of technology development, finding NTC opportunities through vacancies 1, 4 and 5 would be advantageous. Therefore, criticality analysis to discover technologically valuable vacancies should be complimented by trend analysis to provide information on technological trend information according to the purpose or scope of technology development envisaged by the firm conducting the analysis.

4.3.2. *Feature analysis results*

If the development objective is to create further worthwhile technology developments from existing technologies, vacancy 3 is the most promising of the six vacancies, and is selected for further analysis. This analysis may be performed in various dimensions or variables. Keywords and technical fields may be investigated from the illustration. First, keywords in adjacent patents are listed to provide an outline insight into the characteristics of vacancy 3. Since it is unnecessary to include all the keywords in adjacent patents, only a selective set of representative keywords is picked, which is inevitably a subjective decision. In the current research, a keyword is selected if the average number of its appearances per patent is over five. Seven keywords meet the criterion, and are displayed in Table 8. If users want to see more keywords related to the vacancy, the value can be reduced. On the other hand, if users want to see only core keywords, the value needs to be raised.

Second, adjacent patents are analyzed with respect to basic properties of their technical fields. Table 9 exhibits the basic properties of the 15 adjacent patents, of which four are related to display technology, two to input/output devices, six to communication including telecommunicating and the final three to data processing. Although these 15 patents belong to somewhat different technological fields, they are closely related to each other in terms of their keyword features, which collectively locate them in the same area on the patent map. A promising opportunity for NTC may be excavated by in-depth analysis across individual patents. Clearly, the task necessitates a comprehensive set of domain knowledge. The supports of this approach are limited to providing a list of patents that have a high possibility of NTC and their technology features in common. How to create new technology based on the patents in the list is a role of experts.

Table 8
List of representative keywords.

Keyword	Average appearance
Operation	5.47
Data	44.60
Memory	8.67
Output	5.91
Storage	8.00
Area	13.73
Processor	16.47

Table 9
List of adjacent patents for Vacancy 3.

Patent number	Title
4 083 797	Nematic liquid crystal compositions
4 612 584	Screen gravure engraving system for electromechanical engravers
5 526 481	Display scrolling system for personal digital assistant
5 606 594	Communication accessory and method of telecommunicating for a PDA
5 699 244	Hand-held GUI PDA with GPS/DGPS receiver for collecting agronomic and GPS position data
5 819 227	Tour schedule processor for moving bodies
6 168 331	Case keyboard
6 198 941	Method of operating a portable communication device
6 244 873	Wireless myoelectric control apparatus and methods
6 270 271	Printer for portable information processor
6 305 603	Personal digital assistant-based financial transaction method and system
6 360 172	Generation and distribution of personalized multimedia natural-phenomenological information
6 366 450	Hideaway integrated docking cradled
6 424 369	Electronic device having a display
6 512 515	Data compression through motion and geometric relation estimation functions

4.4. Discussions

The suggested approach is based on the two-dimensional patent mapping, which has several advantages over the basic patent statistics (Englesman and van Raan, 1994). Firstly, it is easy to understand the patent data when it is visualized in two-dimensional space rather than summarized just in tables. Secondly, the overall structure of the data can be investigated in short time and the results shown in the patent map are easy to remember. Finally, the latent meanings in the data are easy to be explored by elimination any noise from raw data. In general, a large scale of data inevitably contains a noise, which needs to be removed to reveal the significant meanings from the data. The keyword-based patent map enables to distinguish only the necessary information from the unnecessary data.

Despite those advantages, two-dimensional mapping may cause an information loss. However, as the ultimate goal of this research is not to visualize the patent data but to explore technological vacancies, there is less need to put all knowledge extracted from the patent documents exactly on a two-dimensional space. On the contrary, considering that this research is to discover "possibilities" of NSC, limiting its focus only to the main information showing a significant variance of keyword frequency would be a good strategy. Actually, various possibilities for NSC could be explored from the several combinations of two PCs out of several PCs extracted from the PCA results.

To apply the suggested approach, a plenty of time and efforts may be required in developing the suggested map and analyzing the information on the map. More efficient way to save time and efforts is required to increase the utility of the approach. For the purpose, we designed a supporting system to deal with a large amount of patent data, which was incorporated in a knowledge management system (KMS) as its sub-function.[1] Without the system, all

[1] The suggested system has also been integrated into a KMS developed by Seoul National University over the past 5 years to address 'KMS for 4th generation R&D', called KNOWVATION (Park and Kim, 2006). It divides R&D activities into three stages: strategy, execution and evaluation. Within this structure, the suggested system for NTC functions as a sub-module of the first stage, with the aim of identifying potential R&D programs. The system allows a keyword-based patent map for a specific technology field to be developed with little human effort. When possible patent vacancies have been roughly defined by expert judgment, the system automatically identifies adjacent patents for each vacancy and provides vacancy validity test results, which should be reviewed by experts.

the steps proposed in this research should be done manually and thus time consuming. However, once the system is developed, the mapping and test results can be obtained automatically, saving time and cost significantly. Of course, user inputs are necessary to set criteria to get those results. For example, candidates for keyword list can be generated automatically but the final list should be determined by users. Similarly, candidates for adjacent patents can be obtained automatically, but the final patents for subsequent analysis should be selected by users.

Therefore, it may be possible to reduce experts' manual work by developing a supporting system, and yet the analysis and interpretation of experts having domain knowledge is indispensable not only during the selection process of keywords or adjacent patents but also throughout the whole process of this research. In most data-mining applications, data-mining techniques do not eliminate the need for human input (Piramuthu, 2004). They indeed requires experts to set the boundaries of the analysis (e.g. to provide query to gather patent documents of concern), a process known as feature selection (e.g. to determine a keyword list, vacancy boundaries, or adjacent patents), and interpret the results of the analysis (e.g. to decide finally the vacancies to be examined in detail). These are particularly important when applying data-mining to patent analysis, due to the complex linguistics embedded in patent files (Fattori et al., 2003). Though large parts of patent analysis can be automated, the role of experts must be emphasized.

5. Conclusions

This article presents a new approach for developing keyword-based patent maps and applying them to the idea generation phase of NTC practice. Compared to a conventional bibliographic patent map, the keyword-based patent map has considerable advantages in terms of information extraction, visualization and analysis. Operational efficiency is also enhanced, as the keyword-based patent map reduces the burden of manual work.

The application of patent mapping to NTC practice represents an unprecedented experiment, and thus accounts for the major contribution of the current research. The underlying idea is to excavate the latent characteristics of patents from a patent map and identify some unexplored vacancies in the map. These are blank zones which are surrounded by numerous existing patents, and which can be expected to provide potential for future NTC. Promising NTC opportunities can be derived by intensively examining surrounding patents. Above all, the focus of this research is not limited to the development of the patent map. Rather, this research emphasizes on how to analyze, interpret and utilize the patent map to discover new technology opportunities that may have NTC potential, and provide an algorithm for examining this potential. All activities in the process that could be computerized have been systemized, which saves considerable time and effort in generating the map, thus giving specific practical help to staff in charge of NTC. Also, the suggested approach can be applied to business model patents in order to discover new business opportunities. The same procedures can be used in identifying new business creation potential instead of NTC potential. Although more business models and software solutions are patented, they haven't been yet analyzed actively. Those patents can be good sources of new business creation and should be addressed in the future research.

By its nature, this study is an exploratory one, and needs more extension and/or elaboration in terms of methodology and application. Ensuing research may consider the following issues among others. First, the validation of vacancy importance should be extended. For instance, market analysis needs to be incorporated, since technical progress and customer needs may provide richer information for the process. Second, the whole process needs to be systemized and automated. Although we have developed an expert system and automated some elements, there is still considerable scope for further work to enhance operational efficiency. Third, the validity of this approach necessitates more testing work by employing patent documents from a wider range of technologies, which is indispensable for gaining external validity. In addition, real case studies in the company setting will be required in the future and we are planning to continue the research. Fourth, more techniques for patent mapping, especially putting the information on a two-dimensional space, need to be investigated. Though PCA was adopted in this study, which was regarded as the most suitable technique, it must be meaningful to compare the results of PCA with those of SOFM, CA, and possible other mapping techniques. Actually, PCA is subjected to the information loss that might be serious in other applications. Thorough review of more techniques will be helpful for increasing the effectiveness of this approach. Finally, but most importantly, the current research must be elaborated to identify specific meanings of PCs and extended to develop more concrete NTC specifications. The outcome of the current research may generate a set of patents that surely contribute to reducing the domain for exploration, but without actually being able to pinpoint the exact specification of new product proposal: this is a task for the future.

Appendix A. Comparisons of 'all-keyword approach' and 'major-keyword approach'

All-keyword approach utilizes all keywords in constructing the keyword vector. Major-keyword approach, on the other hand, uses only a selective set of the most significant keywords. The former may reduce information loss, but the latter may improve the explanatory power in PCA results. Major-keyword approach again can be divided into two types according to the way of selecting major keywords. Major keywords can be defined as the keywords that show a high frequency of appearance in the documents on the assumption that more important keywords are

Table A1
Comparisons of approaches in terms of explanation power.

Comparison	All-keyword approach (A)	Major-keyword approach	
		High frequency (B)	High variance (C)
Number of keywords	39	11	12
Explanation power of 4 PCs (%)	72	62	39

Table A2
Comparisons of approaches in terms of average difference of PC 1.

Comparison	A–B	B–C	A–C
Average difference of PC 1	0.16	0.09	0.20

mentioned more frequently. Or they also can be defined as the keywords that show a high variance of appearance among the documents because such keywords represent the technological features that distinguish patent documents quite well.

For the purpose of comparisons between three approaches, we conducted an experimental test in terms of explanation power and average difference of the first PCs. Here, the whole set of 39 keywords was used for the all-keyword approach (A), but only the 11 keywords with the highest frequency were applied for the first major-keyword approach (B) and the 12 keywords with the highest variance were used for the second major-keyword approach (C). The comparison results are provided in Tables A1 and A2. Table A1 shows that the four PCs extracted from A can still explain 72% of the sample variance, those from B can guarantee 62%, and finally those from C can preserve only 39%, which signifies that the information loss might be serious in C. As shown in Table A2, however, the differences between approaches in terms of average difference of PC1 are negligible, ranging from 0.09 to 0.20. It implies that the mapping structure compared to the information loss is quite robust to the number of keywords. And thus, all-keyword approach is adopted in this research, which is similar to other approaches in its mapping structure but can keep much more information than others.

References

Abraham, B., Morita, S., 2001. Innovation assessment through patent analysis. Technovation 21, 245–252.

Andal, M., Oyanagi, S., Yamakazi, K., 2006. Research on text mining techniques to support patent map generation. Forum on Information Technology, 111–112.

Ashton, W., Sen, R., 1988. Using patent information in technology business planning—II. Research Technology Management 32, 36–42.

Austin, D., 1993. An event-study approach to measuring innovative output: the case of biotechnology. American Economic Review 83, 253–258.

Bader, M., 2008. Managing intellectual property in the financial services industry sector: learning from Swiss Re. Technovation 28, 196–207.

Bay, Y., 2003. Development and applications of patent map in Korean high-tech industry. In: Proceedings of the First Asia-Pacific Conference on Patent Maps, Taipei, October 29, 2003, pp. 3–23.

Campbell, R.S., 1983. Patent trends as a technological forecasting tool. World Patent Information 5 (3), 137–143.

Camus, C., Brancaleon, R., 2003. Intellectual assets management: from patents to knowledge. World Patent Information 25 (2), 155–159.

Chesbrough, H., 2003. Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business School Press, Boston.

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

Davidson, G., Hendrickson, B., Johnson, D., Meyers, J., Wylie, B., 1998. Knowledge mining with VxInsight: discovery through interaction. Journal of Intelligent Information System 11, 259–279.

Deerwester, S., Dumais, S., Furnas, G., 1990. Indexing by latent semantic analysis. Journal of American Society for Information Science 41 (6), 391–407.

Delegation of Japan, 2000. OWAKE system—primary automatic classification. WIPO Report 59, pp. 13–17. Available at 〈http://www.wipo.int/classifications〉.

Dolcera website. Available at 〈http://www.dolcera.com〉.

Englesman, E.C., van Raan, A.F.J., 1994. A patent-based cartography of technology. Research Policy 23, 1–26.

Ernst, H., 1995. Patenting strategies in the German mechanical engineering and their relationship to company performance. Technovation 15, 225–240.

Ernst, H., 2001. Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. Research Policy 30, 143–157.

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

Fattori, M., Pedrazzi, G., Turra, R., 2003. Text mining applied to patent mapping: a practical business case. World Patent Information 25, 335–342.

Fayyad, U., Piatetsky-Shapiro, P., Smyth, P., Uthurusamy, R., 1996. Advances in Knowledge Discovery and Data Mining. AAI Press, CA.

Fischer, G., Lalyre, N., 2006. Analysis and visualisation with host-based software—the features of STN®AnaVist™. World Patent Information 28 (4), 312–318.

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

Fujitsu website. Available at 〈http://glovia.fujitsu.com〉.

Greenacre, M., 1984. Theory and Applications of Correspondence Analysis. Academic Press, New York.

Hanel, P., 2006. Intellectual property rights business management practices: a survey of literature. Technovation 26 (8), 895–931.

Hull, D., Aït-Mokhtar, S., Chuat, M., Eisele, A., Gaussier, É., Grefenstette, G., 2001. Language technologies and patent search and classification. World Patent Information 23, 265–268.

Invengine website. Available at 〈http://www.invengine.net〉.

Japan Institute of Invention and Innovation (JIII), 2002. Guide Book for Practical Use of Patent Map for Each Technology Field.

Johnson, R., Wichern, D., 1998. Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliff, NJ.

Jung, S., 2003. Importance of using patent information. In: WIPO—Most Intermediate Training Course on Practical Intellectual Property Issues in Business. World Intellectual Property Organization (WIPO), Geneva, pp. 10–14.

Karki, M., 1997. Patent citation analysis: a policy analysis tool. World Patent Information 19 (4), 269–272.

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.

Kohonen, T., 1995. Self-organizing Maps. Springer, Berlin.

Kostoff, R., Toothman, D., Eberhart, H., Humenik, J., 2001. Text mining using database tomography and bibliometrics: a review. Technological Forecasting and Social Change 68, 223–252.

Krier, M., Zaccà, F., 2002. Automatic categorisation applications at the European patent office. World Patent Information 24, 187–196.

Kuznets, S., 1962. Innovative activity: problems of definition and measurement. In: Nelson, R. (Ed.), The Rate and Direction of Inventive Activity. Princeton University Press, New Jersey.

Lanjouw, J., Schankerman, M., 1999. The quality of ideas: measuring innovation with multiple indicators. National Bureau of Economic Research, 7345.

Larkey, L., 1999. A patent search and classification system. In: Proceedings of the Fourth ACM Conference, 1999, pp. 179–187.

Leclercq, I., 1999. INPI, the Internet and electronic commerce. World Patent Information 21, 259–265.

Lee, S., Lee, S., Seol, H., Park, Y., 2008. Using patent information for designing new product and technology: keyword-based technology roadmapping. R&D Management 38 (2), 166–188.

Lerner, J., 1994. The importance of patent scope: an empirical analysis. RAND Journal of Economics 25, 319–332.

Liu, S., Shyu, J., 1997. Strategic planning for technology development with patent analysis. International Journal of Technology Management 13 (May), 661–680.

Liu, S.J., 2003. A route to a strategic intelligence of industrial competitiveness. In: Proceedings of the First Asia-Pacific Conference on Patent Maps, 2003, pp. 2–13.

Lyon, M., 1999. Language related problems in the IPC and search systems using natural language. World Patent Information 21, 89–95.

M-Cam website. Available at ⟨http://www.m-cam.com⟩.

Miller, W., Morris, L., 1999. 4th Generation R&D: Managing Knowledge, Technology, and Innovation. Wiley, New York.

Morris, A., Wu, Z., Yen, G., 2001. A SOM mapping technique for visualizing documents in a database. In: Proceedings of the IEEE International Joint Conference on Neural Networks, Washington DC, USA, July 2001.

Morris, S., DeYong, C., Wu, Z., Salman, S., Yemenu, D., 2002. DIVA: a visualization system for exploring documents databases for technology forecasting. Computers & Industrial Engineering 43 (4), 841–862.

Neopatents website. Available at ⟨http://www.neopatents.com⟩.

Norton, M.J., 2001. Introductory Concepts in Information Science. ASIA, New Jersey.

Park, Y., Kim, S., 2006. Knowledge management system for fourth generation R&D. Technovation 26 (5–6), 595–602.

Park, Y., Yoon, B., Lee, S., 2005. The idiosyncrasy and dynamism of technological innovation across industries: patent citation analysis. Technology in Society 27 (4), 471–485.

Piramuthu, S., 2004. Evaluating feature selection methods for learning in data mining applications. European Journal of Operations Research 156, 483–494.

Salamatov, Y., 1999. TRIZ: The Right Solution at the Right Time—A Guide to Innovative Problem Solving. Insytec, Hattem.

Salton, G., Buckley, C., 1998. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5), 513–523.

Schellner, I., 2002. Japanese file index classification and F-terms. World Patent Information 24, 197–201.

Scherer, F.M., 1981. Using linked patent and R&D data to measure inter-industry technology flows. In: Griliches, Z. (Ed.), R&D, Patents, and Productivity. University of Chicago Press for NBER, Chicago.

Schűtze, H., Pedersen, J., 1994. A co-occurrence-based thesaurus and two applications to information retrieval. In: Proceedings of the RIAO '94 Conference.

Schűtze, H., Pedersen, J., 1995. Information retrieval based on word sense. In: Proceedings of Fourth Annual Symposium, 1995, pp. 161–176.

Shane, S., 2001. Technological opportunities and new firm creation. Management Science 47, 205–220.

Slater, S., 1993. Competing in high-velocity markets. Industrial Marketing Management 22 (1), 225–263.

Smith, H., 2002. Automation of patent classification. World Patent Information 24, 269–271.

Takeuchi, H., Nonaka, I., 1986. The new product development game. Harvard Business Review 64, 137–146.

Theodorou, Y., Drossosb, C., Alevizos, P., 2007. Correspondence analysis with fuzzy data: the fuzzy eigenvalue problem. Fuzzy Sets and Systems 158, 704–721.

Thomson website. Available at ⟨http://scientific.thomson.com/products/aureka⟩.

Tseng, Y., Juang, D., Wang, Y., Lin, C., 2005. Text mining for patent map analysis. In: Proceedings of IACIS Pacific 2005 Conference, May 19–21, Taipei, Taiwan, 2005, pp. 1109–1116.

Tseng, Y., Lin, C., Lin, Y., 2007a. Text mining techniques for patent analysis. Information Processing and Management 43 (5), 1216–1247.

Tseng, Y., Wang, Y., Lin, Y., Lin, C., Juang, D., 2007b. Patent surrogate extraction and evaluation in the context of patent mapping. Journal of Information Science 33 (6), 718–736.

Uchida, H., Mano, A., Yukawa, T., 2004. Patent map generation using concept-based vector space model. In: Proceedings of the Fourth NTCIR workshop, June 2–4, Tokyo, Japan.

Urban, G., Hauser, J., 1993. Design and Marketing of New Products. Prentice-Hall, Englewood Cliffs, NJ.

USPTO website. Available at: ⟨http://www.uspto.gov⟩.

Utterback, J., 1996. Mastering the Dynamics of Innovation. Harvard Business School Press, Massachusetts.

von Hippel, E., 1986. Lead user: a source of novel product concepts. Management Science 32, 791–805.

VxInsight website. Available at ⟨http://www.metricsgroup.com⟩.

Wang, P., Cockburn, I., Puterman, M., 1998. Analysis of patent data—a mixed Poisson-regression-model approach. Journal of Business & Economic Statistics 16, 27–41.

Watanabe, C., Tsuji, Y., Griffy-Brown, C., 2001. Patent statistics: deciphering a 'real' versus a 'pseudo' proxy of innovation. Technovation 21 (12), 783–790.

Weiss, S., Indurkhya, N., Zhang, T., Damerau, F., 2005. Text Mining Predictive Methods for Analyzing Unstructured Information. Springer, Berlin.

WIPO, 2003. Patent map with exercises (related). WIPO-MOST intermediate training course on practical intellectual property issues in business, Theme 16.

Yeap, T., Loo, G., Pang, S., 2003. Computational patent mapping: intelligent agents for nanotechnology. In: Proceedings of the International Conference on MEMS, NANO and Smart Systems, 20–23 July 2003, 2003, pp. 274–278.

Yoon, B., 2005. Methodology for managing technological knowledge and developing new technology using patent analysis. Ph.D. Thesis, Seoul National University.

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

Yoon, B., Phaal, R., Probert, D., 2008. Structuring technological information for technology roadmapping: data mining approach. In: Proceedings of the WSEA Conference, Cambridge, UK, February 2008.

Yoon, B., Yoon, C., Park, Y., 2002. On the development and application of a self-organizing feature map-based patent map. R&D Management 32 (4), 291–300.