Scientometrics (2010) 85:803–820
DOI 10.1007/s11192-010-0294-5

Development and application of a keyword-based knowledge map for effective R&D planning

Byungun Yoon · Sungjoo Lee · Gwanghee Lee

Received: 24 January 2010 / Published online: 23 September 2010

Abstract With the growing recognition of the importance of knowledge creation, knowledge maps are being regarded as a critical tool for successful knowledge management. However, the various methods of developing knowledge maps mostly depend on unsystematic processes and the judgment of domain experts with a wide range of untapped information. Thus, this research aims to propose a new approach to generate knowledge maps by mining document databases that have hardly been examined, thereby enabling an automatic development process and the extraction of significant implications from the maps. To this end, the accepted research proposal database of the Korea Research Foundation (KRF), which includes a huge knowledge repository of research, is investigated for inducing a keyword-based knowledge map. During the developmental process, text mining plays an important role in extracting meaningful information from documents, and network analysis is applied to visualize the relations between research categories and measure the value of network indices. Five types of knowledge maps (core R&D map, R&D trend map, R&D concentration map, R&D relation map, and R&D cluster map) are developed to explore the main research themes, monitor research trends, discover relations between R&D areas, regions, and universities, and derive clusters of research categories. The results can be used to establish a policy to support promising R&D areas and devise a long-term research plan.

S. Lee (✉)
Department of Industrial & Information Systems Engineering, Ajou University,
San 5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, South Korea
e-mail: sungjoo@ajou.ac.kr

B. Yoon
Department of Industrial & Systems Engineering, Dongguk University-Seoul,
26, Pil-dong 3 ga, Jung-gu, Seoul 100-715, South Korea
e-mail: postman3@dongguk.edu

G. Lee
Public Administration, National Research Foundation of Korea,
180-1, Gajeong-dong, Yuseong-gu, Daejeon 305-350, South Korea
e-mail: thomas@nrf.go.kr

Keywords Knowledge map · Research proposal database · Text mining · Network analysis · R&D

Introduction

The evaluation of public research and development (R&D) programs has been highlighted since the 1980s when the concept of strategic research management was introduced (Hayashi 2003). Since R&D programs can facilitate individual projects, their evaluation is an important aspect of management in a research and innovation system. In addition, scientific collaboration is also a critical issue in scientific activity because open innovation becomes one of the widespread success factors to realize technological and economical benefits. In the field of bibliometrics, the evaluation of R&D programs and scientific collaboration has been a focal theme of research because a wide range of documents can be analyzed to investigate the trends in R&D, as well as the relationships between companies and researchers in R&D networks (Moed et al. 1991). Bibliometric assessments are economical, non-invasive, and simple to implement, permitting updates and rapid inter-temporal comparison with more quantitative data (Abramo et al. 2009). Thus, a wider use of bibliometrics to evaluate the quality and efficiency of research activities is realized in areas of scientific investigation that are well represented by articles in international journals (Van Raan 2005). The rapid growth in science and technology enhances the popularity of bibliometrics by causing an exponential increase in the number of academic papers (more than 500,000 per year) and patents (Moon 2005). However, while greater accessibility and a large range of databases have enhanced the usability of available information, little attention has been paid to screening, classifying, and interpreting information (Yoon 2008). In order to stimulate innovative activities at both the national and firm levels, the accumulated information on science and technology should be analyzed and mined to discern meaningful patterns in such activities. As a result, refined knowledge can support a decision-making process for conducting R&D projects, thereby leading to the discovery of promising research themes.

In general, information on science and technology, such as academic research, has been disseminated in the form of patents or papers. Therefore, a process for generating scientific and technological knowledge is dependent mainly on mapping meaningful knowledge that is extracted from quantitative data with regard to patents and papers. Knowledge maps can be defined as a visual form to help grasp the characteristics and implications of a large amount of information by analyzing the types of information and patterns. Basically, such practical needs are overwhelming in R&D management because many patents and papers (unstructured data) need to be transformed into well-structured data, thereby allowing researchers to apprehend the historical trends in research and plan long-term R&D projects. Although useful information, such as keywords, citations, and co-authoring, can be examined, the level and the quality of mapping have still remained primitive. Thus, in this paper, a knowledge map for R&D refers to a map that captures the overall trend in research activities by analyzing comprehensive information that is obtained from R&D proposal databases.

Many researchers have proposed the concept of knowledge maps and a variety of methods for drawing them. In R&D management, knowledge maps have been developed to implement a policy in relation to budgeting for national research agendas or devise a strategy for acquiring new technology. However, traditional approaches for generating knowledge maps are subject to limitations in various aspects. First, the existing processes for developing knowledge maps are dependent mostly on the synthesized opinions of experienced experts, resulting in inefficiency and possibly low quality. The adoption of a bibliometric approach will be an appropriate remedy for this limitation because a large amount of data can be analyzed in a systematic manner, covering a broad spectrum of information on a specific subject of interest. Second, traditional approaches that develop knowledge maps by applying bibliometric information have a tendency to utilize citation relationship or predefined keywords. However, since the approaches are subject to the size or depth of personal knowledge or network of authors, new methods to develop a knowledge map need to analyze the contents of documents in order to reflect real relationship among papers, research proposals and patents. Text mining can be a potential technique to generate a knowledge map on the basis of content analysis. Third, most of the existing knowledge maps concentrate on the visualization of a static relationship among researchers or knowledge itself. Thus, the trends of visual forms that can present dynamic changes of research activities have been hardly investigated. In particular, despite their important roles in providing meaningful implications, valuable indices have scarcely received attention from researchers. Fourth, while patent and paper databases are actively analyzed, research proposal databases have been underutilized despite their potentially high value. Since most research proposals generally consist of natural-language forms, the handling of unstructured data is demanding in terms of effort and cost. Finally, conventional research has failed to propose sophisticated knowledge maps that are dedicated to R&D and can provide valuable information on research activities. A considerable body of literature emphasizes the strategic and practical use of knowledge maps. Thus, concrete processes and methodologies should be proposed and implemented in practical projects for drawing various types of knowledge maps that would help researchers anticipate promising R&D subjects from research proposal databases through bibliometric analysis.

This paper aims to suggest a method for drawing a keyword-based knowledge map that uses the research proposal database of the Korea Research Foundation (KRF), which supports the research of academicians through funding from the Korean government. Since the database consists of a wide range of documents, keyword-based knowledge maps—a newly proposed methodology—will be developed through a text mining technique that extracts keywords from documents and characterizes research through keywords. In particular, various types of knowledge maps are designed to elicit strategic implications for selecting promising research projects and supporting them from a funding perspective.

The paper is organized as follows. In “Background” section, a literature survey of knowledge maps, bibliometric analysis, and text mining is presented to clarify the relevant theoretical and practical background. Then, in “Keyword-based knowledge map” section, the research framework of this paper is explained in terms of the concept, process, and methodology for analysis. In “Illustrations” section, the results of the analysis are illustrated to depict how the proposed approach can be implemented; this is accomplished by presenting illustrative knowledge maps and indicators. Finally, in “Conclusions and future research” section the implications and limitations of current research and issues for further research are discussed.

Background

Knowledge maps

Undoubtedly, useful tools for facilitating knowledge activities will affect the possibility of success in the implementation of knowledge management systems (KMS). Many areas of research aim to advance the understanding of concept formation and utilization in perception and improve ontology design, optimization, and usage for knowledge-sharing activities (Goldstone and Kersten 2003; Gruber 1995). In particular, concept maps that were proposed by Novak add a visual means to communicate knowledge structures for sharing and consensus finding (Novak 1998). A knowledge map is a representative tool or technique that enables visualizing knowledge and relationships in a clear form so that the relevant features of the knowledge can be precisely highlighted (Vail 1999). It is a method designed not only to elicit the knowledge that a decision maker faces, but also to combine probabilities associated with various factors in order to obtain a final probability (Browne et al. 1997). Therefore, knowledge maps can play a key role in creating a more effective KMS; developing advanced knowledge maps has become an important issue in the knowledge management domain (Herl et al. 1999; Maule 1997). Traditional knowledge maps usually adopt mathematical and statistical methods to generate maps on the basis of papers, textbooks, and reports (Pritchard 1969). However, the range of data for mapping has been extended to citations, authors, titles, indices, and so on. The most recent trend in knowledge mapping is the application of text mining, which allows automatic tagging and various types of mapping. Thus, co-word maps, citation maps, and co-authorship maps have been developed to enhance the usefulness of knowledge maps.

The concept of knowledge maps has been suggested to provide the “big picture” of science; this emphasizes the importance of computers for constructing multi-dimensional maps (Doyle 1961). A typology for demonstrating the intellectual structure of science has been devised to describe patterns of scientific papers (Price 1965) and track fast-changing research areas that are the focus of special funding efforts by using scientometric journal mapping (Leydesdorff et al. 1994). In particular, a science map that covers all subject areas of science has been generated on the basis of citation information (Griffith et al. 1974) and at the level of the Institute for Scientific Information (ISI) categories using the aggregated journal–journal citation matrix (Leydesdorff and Rafols 2009).

Citation network analysis is a powerful tool to examine the process of creation and transfer of knowledge through scientific publications, and a bibliometric map can identify influential terms and theories associated with the terms (Calero-Medina and Noyons 2008). In addition, the University of California San Diego (UCSD) Basemap of Science was drawn based on 7.2 million papers and over 16,000 journals and proceedings from Thomson Reuters’ Web of Science and Elsevier’s Scopus database (Klavans and Boyack 2007) and a patent map that investigates the relationship between science and technology has been intensively studied (Boyack and Klavans 2008).

While the above studies dealt with a network of subjects, other papers tackled clusters of research areas. The topic clusters were created by applying a co-word analysis to the keywords in the citing publications (Noyons 1999) and a super cluster of the biomedical area was proposed to explain the evolution of biomedical technology and fuse several subjects pertaining to technology (Aaronson 1975).

Many studies visualize bibliometric information to evaluate a geographic region’s performance in a research field (Noyons et al. 1998) and grasp the ontology and taxonomy of knowledge (Borner et al. 2007). In particular, Garfield applied multi-dimensional scaling (MDS) in order to group various research areas into clusters, thereby producing an atlas by analyzing biochemistry and molecular biology (Garfield 1963). White and his colleagues showed that a self-organizing map (SOM) and MDS are able to yield useful outputs more effectively than traditional methods (White et al. 1998). Author maps were proposed to clarify the relationships between authors through co-author analysis (Chen and Paul 2001) or citation networks (Small 1977). Authorlink was implemented on the basis of

Pathfinder networks and CAMEOs, providing information on co-citations (White et al. 2000; Chen 1999; White 2001).

Although the use of knowledge maps varies according to their objectives, their main application can be summarized as follows. First of all, knowledge maps can be used to search for potential domain knowledge in a specific research area. They are very useful in monitoring a tipping point or paradigm for conducting research and forecasting new technology. They also can be applied to visualize the current status of research fusion and integration across technologies and industries (STEPI 2003). The National Institute of Science and Technology Policy (NISTEP) of Japan developed three types of science maps (relation map, individual research area map, and correlation map) by analyzing scientific information extracted from patent and journal databases (NISTEP 2007).

Bibliometric analysis and text mining

The approaches currently in use for R&D planning can be assigned to two categories: peer review and bibliometric techniques (Abramo et al. 2008). While in peer review, judgment is entrusted to a panel of experts that synthesizes a judgment based on the examination and appraisal of parameters, such as quality and socioeconomic impact; bibliometric techniques are based on indicators that are elaborated from data that can be found in publication databases. Recently, some researchers have demonstrated that bibliometric techniques provide useful information that can counterbalance shortcomings and mistakes in peer judgment, such as distortions arising from subjectivity in assessments (Aksnes and Taxt 2004). In addition, the techniques prove to be efficient at guaranteeing low direct costs and notable savings with regard to time, enabling rapid updating and inter-temporal comparison (Narin and Hamilton 1996). In various studies, bibliometric data have been applied to assess research performance and anticipate promising research areas, particularly for the natural and life sciences, because scientific progress is generally achieved by researchers who study research topics by building upon the work of other scientists (Narin 1976; Van Raan 1996). Considerable research has applied bibliometric analysis to assess the scientific basis of research (Van Raan and Van Leeuwen 2002) and investigate trends in R&D activities in specific industries such as biotechnology and nanotechnology (Perry and Rice 1998; Murray 2002; Yoshiyuki et al. 2009).

Among other methods, text mining is regarded as a novel and exceptional technique for enhancing the applicability of bibliometric analysis. Text mining can be defined as a process to automatically extract information and discover patterns from documents (Dixon 1997). Since documents in a textual format have no structured data fields, it is very difficult to efficiently analyze such textual information. Thus, the first step of text mining is to define key features of documents in order to transform unstructured textual information into structured information. In short, text mining puts a set of labels on each document; further, discovery operations are performed using the labels. The technique is particularly notable because it can cluster similar documents (Dhillon and Modha 2001) or classify newly generated documents into a category (Sebastiani 2002). A more intelligent application of text mining involves identifying the subjects of documents and summarizing their content (Clifton 2004). The scope of subjects in text mining is very broad. First, documents are automatically classified into a predefined class. For this, k-nearest neighbor (k-NN) techniques are applied to text classification (Cohen and Hirsh 1998; Yang 1994) and machine learning is utilized for the automatic classification of documents (Sebastiani 2002). Furthermore, a corpus-based method applies text mining techniques on a corpus of web pages to automatically create web directories and organize them into hierarchies

(Yang and Lee 2004). Second, documents are grouped into several clusters on the basis of the similarity of content. K-means and hierarchical clustering algorithms are adopted for such clustering (Dhillon and Modha 2001). While such text clustering algorithms utilize the vector space model, recent methods are proposed to apply the sequences of frequent words or meaning (Li et al. 2008). In particular, SOMs are adopted to improve the performance of the process (Kohonen 1995) and a graph-based approach to document classification allows for a much more expressive document encoding than the more standard bag-of-words approach (Jiang et al. 2010). Finally, the structuring of documents in an unstructured format is a recent research trend that employs mathematical models of machine learning (Hsu and Chang 1999). Automatic tools for assisting engineers or decision makers are suggested to facilitate text segmentation, summary extraction, feature selection, cluster generation, and topic identification (Tseng et al. 2007). Furthermore, a method to extract data from a web page in HTML has been studied in the web-based environment (Hammer et al. 1997) and online forums hotspot detection and forecast using text mining approaches are conducted by automatically analyzing the emotional polarity of a text (Li and Wu 2010). The technique is applied to not only database management and the Web environment but also foresight exercises and the automated curation of scientific literature through biomedical text mining (Santo et al. 2006; Lourenco et al. 2009).

Keyword-based knowledge map

Concepts and types

The database of the National Research Foundation (NRF), which is chosen as the main database in this paper, is composed of a large number of research proposals that are submitted to its academic research support program. The research proposals include abundant information on the objectives, background, research ideas, and references of the proposed studies. Although the documents include a large amount of content that is dedicated to technology analysis, it is very difficult to investigate all the information of a document. Thus, quick navigation and summary of documents is critical to successful analysis in R&D management. The keywords that are extracted from content in documents are a powerful vehicle to efficiently manage a given amount of information and be able to translate sophisticated research ideas into real outputs of R&D. Therefore, the development of knowledge maps based on keywords is influential because critical R&D databases that have not been analyzed can be applied to generate meaningful knowledge maps; moreover, the process of developing knowledge maps can be automatically accomplished with minimal manual intervention. While traditional knowledge maps focus on gathering the domain knowledge of experts, keyword-based knowledge maps adopt a systematic approach that visualizes the information in a computerized manner through text mining. To this end, a document needs to be transformed into keyword vectors that are composed of the frequency of occurrence of each keyword in the document. The keyword vectors of documents are applied to examine the trends, relationships, and clusters of technologies through graphs, statistical analysis, and indices.

From the basic concepts, five types of knowledge maps are developed in this research: core R&D maps, R&D concentration maps, R&D trend maps, R&D relation maps, and R&D cluster maps. The objective of a core R&D map is to derive state-of-the-art R&D subjects by visualizing the importance of the subjects. R&D concentration maps aim at examining the concentration ratio of specific R&D areas through Herfindahl's index (HI).

R&D trend maps present emerging/declining R&D subjects. R&D relation maps and cluster maps simultaneously can offer concrete links as well as groups featuring similar R&D activities. Since R&D knowledge has not been effectively managed and there are few methods for supporting technology management activities, a keyword-based knowledge map has great potential to provide intelligent knowledge for R&D planning and valuation.

Process

As mentioned above, this research aims at presenting a new approach to create five types of knowledge maps by visualizing information in the research proposal-document database. Thus, a process for drawing a keyword-based knowledge map should be implemented in a stepwise manner as follows. First, the proposal documents of interest need to be collected from the database. Second, since the format of documents is unstructured without data fields, the data therein should be transformed into structured data (keyword vectors). For this, keywords are extracted by calculating the occurrence frequency of words and excluding peripheral words such as conjunctions and articles. The keyword vector of each document is composed of binary value of extracted keywords. Third, keyword vectors of proposals which are classified into a research category are aggregated to scrutinize relationships across R&D areas by summing the values of data fields (keywords). If the number of extracted keywords is n, keyword vector has n dimensions. The aggregated keyword vector also has n dimensions and the kth data field of it has the sum of values in the kth data field of all keyword vectors. Fourth, the proposed knowledge maps are generated to help researchers investigate promising R&D subjects, trends, relationships between R&D activities, and the similarity of R&D subjects. Thereafter, the interpretations from various knowledge maps should be discussed in detail. Figure 1 describes the stepwise process to develop a keyword-based knowledge map.

Development of keyword-based knowledge map

Core R&D map

Promising R&D subjects should be derived to prioritize a lot of research ideas and effectively allocate research budgets. The core R&D map assists policy makers and researchers in

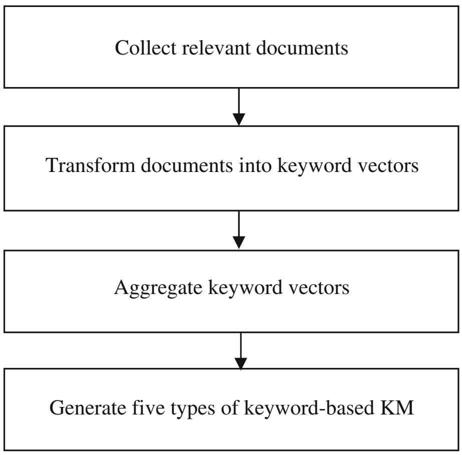

Fig. 1 Process of the proposed approach

understanding the principal research themes and investing research capabilities and R&D funds towards such influential subjects. To this end, important keywords from all the collected documents are extracted through text mining; on the basis of the set of keywords, the documents are transformed into keyword vectors. The keyword vectors should be aggregated according to the predefined research categories to analyze influences and trends of research subjects over research areas. The list of keywords and their occurrence explain which research theme plays a critical role in a given research area, thereby providing definitive information on influential and up-to-date research themes. The core R&D map positions the promising R&D subjects at the central part of the map; in contrast, less important subjects are located at a peripheral part of the map. For this reason, the maximum value of occurrence frequency of keywords is put in the center of the map and the minimum value of it occupies the outside edge of the map. The nearer a R&D theme is positioned to the center of the map, the more crucial the related R&D projects generally are. In addition, the core R&D areas of a country need to be compared to those of other countries. Thus, the proposed core R&D map presents the characteristics of research activities by comparing the focuses of both international and domestic research.

R&D trend map

The historical trends of research are examined to enable researchers to forecast and prepare promising future research themes. While the core R&D map focuses on the currently important research subjects, the R&D trend map presents emerging and declining keywords by examining the frequency of each keyword over time. Since a keyword vector of a paper consists of binary value of keyword occurrence, the number of a keyword in aggregated keyword vectors indicates the number of papers in a research theme. Basically, this process is based on the change in frequency and growth rate of keywords and produces a list of the main keywords and their changing patterns with regard to the frequency. This knowledge map visualizes the fluctuation of the frequency of occurrence of all keywords on a two-dimensional space that presents the growth rate of each keyword as well. Moreover, the trends of research activities of a country are investigated by comparing them with international trends.

R&D concentration map

If researchers in a research area have studied just a few research subjects intensively, a broad spectrum of research subjects cannot be covered. The R&D concentration map applies the HI to calculate the degree of concentration of research areas. The index can be calculated by an equation as follows.

$$HI = \sum S_i^2$$

Here, S_i denotes the ratio for the ith subject that the number of the subject is divided by the total number of all subjects and HI can be calculated by summing the squares of the S_i values. Thus, if subjects have high ratios such as 0.9 or 0.8, HI obviously approaches the upper limit, indicating that in this case, the research areas are concentrated with several research subjects. Since the index is generally an indicator for investigating the level of competition across subjects, it can be used to measure the extent of research concentration. Table 1 depicts the concentration level of a research subject-area based on HI.

Table 1 Level of concentration for various values of the HI

Type	Range of HI	Level of concentration
Perfect competition	0–0.2	Very high
Monopolistic competition	Around 0.2	Medium–high
Oligopoly	0.2–0.7	Low–medium
Monopoly	0.7–1	Very low

R&D relation map

R&D relation maps are of three types according to the objects of analysis: research relation map, regional relation map, and institute relation map in a specific area. The maps aim at investigating the similarity of research areas as well as the relationships across regions/universities, thereby catalyzing open innovation through active research collaboration. Since the concept of relationship in the proposed relation maps is based not on information exchange but on the similarity of keywords among research subjects, the maps need a similarity matrix that is composed of the *cosine* values of pairs of research areas (regions and universities in the regional and institute relation maps, respectively) and that should be constructed to generate a network. The similarity across objects is based on the aggregated keyword vectors, which allow the maps to visualize the relationships across the objects. In general, network analysis provides various useful indices such as degree centrality and betweenness centrality. In this paper, degree centrality is adopted to examine which research areas play a central role in the network, and betweenness centrality can be applied to derive intermediary research areas that connect pairs of related research areas.

R&D cluster map

Generally, a classification is irregularly revised by adding a new concept to the original classification, as a result of which the ensuing classification is disorganized and confusing. Thus, the current research classification of the KRF has a drawback in that several subcategories are classified within an incorrect category, which compromises the coherence and consistency of the classification. Therefore, the classification can be modified by grouping similar subcategories based on the degree of similarity. In addition, the process can yield a new classification of R&D whereby related R&D subjects are grouped within a category to enable policy makers to balance the allocation of R&D budgets. For this reason, the keyword vectors of research proposals in subcategories are aggregated, and the similarity matrix among subcategories is constructed through calculating cosine values of them. Then, hierarchical clustering is employed to derive new categories of research by investigating relevant number of clusters. The outputs can be elaborated to modify an existing research classification system. R&D cluster maps visualize the relations between keyword-based clusters and the original classification.

Illustrations

Data

The main data source is the research proposal database of the KRF that contains 225,234 research proposals across 92,701 programs, which cover a specific time period ranging from 2002 to 2006. The classification of R&D in the database consists of eight categories that are divided into subcategories. In order to analyze the documents of the collected proposals, keywords should be available from the documents. The process of extracting keywords can be conducted in two ways: expert-based and computer-based approaches. Basically, text mining plays the role of extracting important keywords from documents by investigating the frequency of occurrence of each word. In this paper, a text-mining tool is applied to derive several significant keywords; then, an individual keyword is added to the whole set of keywords. Ultimately, keyword vectors of the proposal documents are stored in a database by calculating the frequency of occurrence of all keywords. Consequently, this database includes seven important data fields: year, institution, region, category, subcategory, acceptance decision, and keywords. Although all R&D areas can be selected, this study has chosen the 'management' research area to illustrate the proposed approach for developing a keyword-based KM because the R&D area has a large number of proposals; moreover, keywords of the documents are comprehensive and unique enough that the proposed approach seems to be suitable for this area. Thus, 1,920 research proposals related to the 'management' research area are collected for drawing the KM.

Core R&D map

The 'management' category that this study examines consists of 22 subcategories, such as marketing and finance. After the frequency of occurrence of research subjects is calculated, five highly ranked subcategories (management information systems, marketing, finance, human/organization management, and international management) are derived. In this paper, the core R&D map in the subcategory of management information systems (MIS) is illustrated. In Fig. 2, ten keywords of importance are positioned on the core R&D map on the basis of their frequency of occurrence. As a result, the 'technology adoption model,' 'customer relation management,' 'e-commerce,' and 'knowledge management' are decisive research themes. These subjects are critical components in successfully implementing management information systems. However, in the international research community, general research themes such as information technology, information system and knowledge management are conspicuous in the core R&D map. Thus, specific and unique

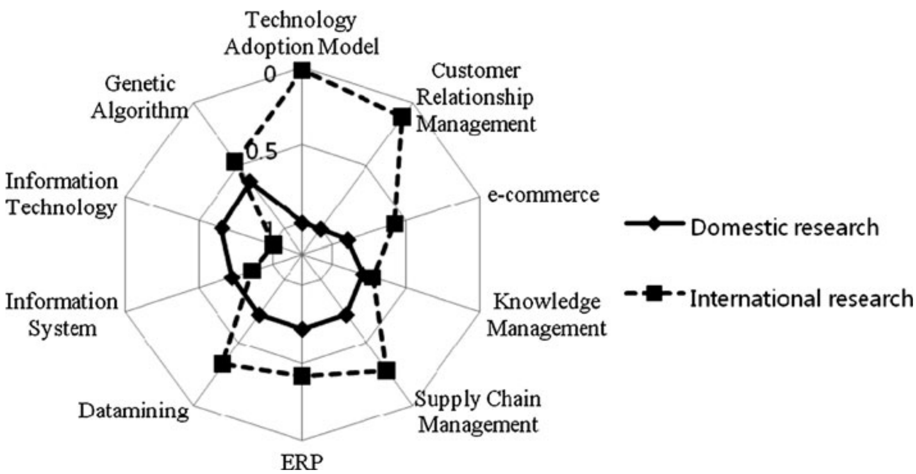

Fig. 2 Core R&D map in the 'MIS' research area

research areas such as ‘technology adoption model’ and ‘customer relation management’ are highlighted in the database of KRF because the Korean government emphasizes the importance of applications and fusion of technology as well as customization of a service.

R&D trend map

The changing patterns of keywords in R&D proposals help researchers to understand trends in research activities and provide information on changes in research subjects. Highly ranked keywords in the MIS category are derived for 2002 through 2006, allowing users to explore emerging and declining keywords. The quantitative criterion to decide whether a keyword is included in the list of emerging or declining keywords is to measure the growth rate of the geometric mean of the keyword occurrence frequency as follows.

$$\text{growth rate} = [\{(1+g_1)(1+g_2)\times\cdots\times(1+g_n)\}^{\frac{1}{n}} - 1]\times 100$$

In this equation, g_i refers to the growth rate from the ith year to the $(i+1)$th year ($n = 4$). While in 2002 and 2003 ‘e-commerce’ was a major topic in the category, the ‘technology adoption model’ has emerged since 2004. In addition, traditional concepts such as ‘B2B’ and ‘information system’ have disappeared over time; in contrast, the ‘technology adoption model’ and ‘ubiquitous computing’ are the focus of the most recent research subjects. Although Fig. 3 visually shows the trends in R&D subjects over a time-horizon, the aforementioned growth rate can present a clear distinction regarding emerging or declining subjects. Thus, in Fig. 3, the labels of the objects in the right-hand side of the graph have the value of the growth rate. Consequently, emerging keywords such as ‘ubiquitous computing,’ ‘expert system,’ and ‘technology adoption model’ reflect the up-to-date trends in research in MIS, whereas research associated with information systems and CRM is traditional or has lost impact in relation to the MIS research agenda.

When such trends of domestic research are compared to those of international research, useful implications of research activities can be derived. International research has a tendency to invest a lot of efforts in ‘ERP’, ‘ubiquitous computing’ and ‘CRM’ shown in

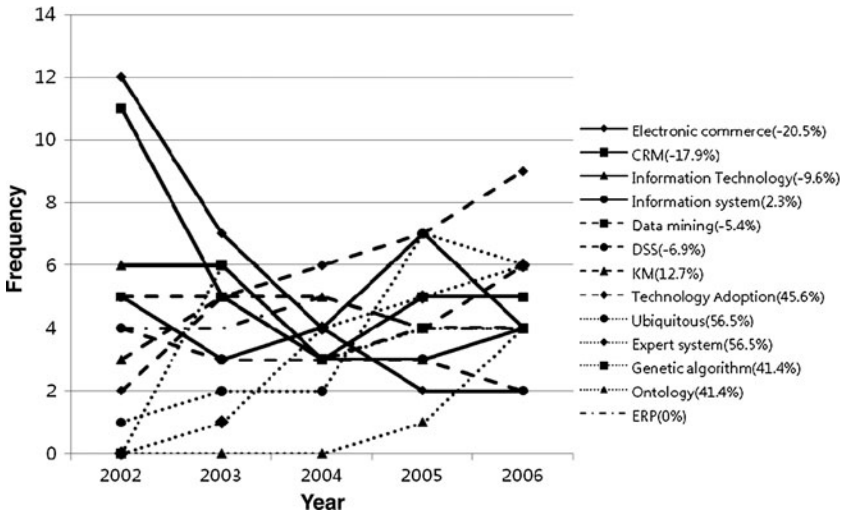

Fig. 3 Trends of keywords in the MIS-related subject (domestic research)

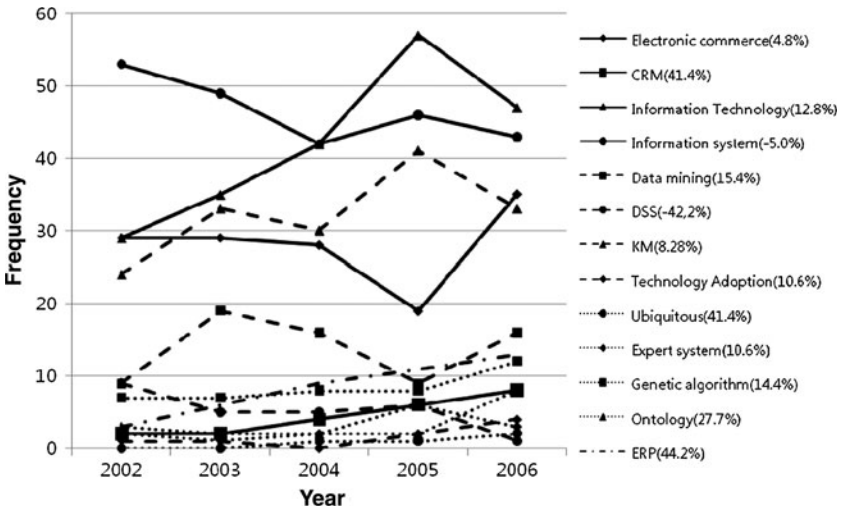

Fig. 4 Trends of keywords in the MIS-related subject (international research)

Fig. 4. While a research area related to ubiquitous computing has been emphasized in both domestic and international research, more academic attention to 'expert system' from the KRF database has been paid than in international research communities. In addition, although 'e-commerce' is regarded as an outdated research area in Korea, many researchers in international research communities still conduct related research.

R&D concentration map

Basically, a critical research avenue can influence a lot of subsequent research in a network of research activities. If only a fraction of the research subjects that are included in a research category are intensively studied, the degree of concentration of the research category is relatively high, indicating that interest in the category is concentrated on only a few subjects. Figure 5 depicts subcategories and the HI values of the various management-related areas. Most research categories are 'oligopolistic' because academics mainly research only a few popular subjects within each of those categories. The MIS research area has the highest value of HI among those categories that are 'oligopolistic.' This category has four subcategories: management computing processes, management information systems, information technology management, and intelligent decision support system (DSS). Among them, information technology management accounts for 64.92%, as a result of which the MIS category can be characterized as an oligopolistic type. Even though most of the categories are characterized as oligopolies, some categories can be characterized as monopolistic competition or perfect competition. In particular, two categories (SCM and field-specific management) manifest rather balanced distributions of subcategories.

From these results, a specific subject in a research category is focused on, indicating that most of the researchers pay attention to only a few themes of the research category. The MIS, general management, production management and sales/marketing categories are oligopolistic categories because a few subcategories of subjects are intensively studied by domain researchers. However, in SCM and field-specific management categories, the interests of researchers are dispersed into a broad spectrum of subcategories because the research themes are unique and valuable to tackle.

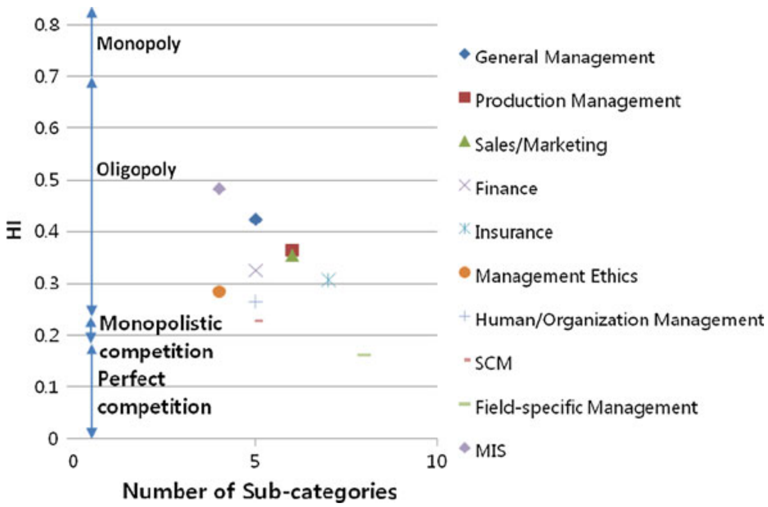

Fig. 5 R&D concentration map in the management-related subject

R&D relation map

The process for developing an R&D relation map is based on the calculation of cosine-based similarities across research categories. A similarity matrix is used to generate a research network that consists of nodes (research categories) and links (relations). If a pair of research categories exhibits more similarity than a cut-off value, the two research categories are linked in a network. Since the cut-off value is critical for drawing an accurate network, sensitivity analysis is executed to derive a proper cut-off value. In this paper, the similarity matrix of 22 sub-categories in the 'management' category is constructed by aggregating the keyword vectors of research proposals in each subcategory. Figure 6 shows the research network of the 'management' category. The 'MIS' and 'strategic management' categories play a central role in the network; on the other hand, the 'technology management' category facilitates links between pairs of categories, which means that the category often resides in the linkages between other categories. This can be clarified by calculating two indices, viz., the degree centrality and the betweenness centrality. In Table 2, the list of highly ranked research areas with regard to degree centrality and betweenness centrality is presented for the management category. The management strategy/policy and MIS categories have high values for the two indices, indicating that these two categories have a lot of relationships with other categories and simultaneously intermediate such relations between pairs of categories. In particular, although the degree centrality of the 'technology management' category is relatively low, its betweenness centrality is highly ranked, showing that the category is a successful intermediary in the research network.

R&D cluster map

The proposed R&D cluster map is useful to understand differences between the existing classification of research and new clusters based on the similarity of keywords by visualizing the relations regarding the classification and clusters. Figure 7 presents an example

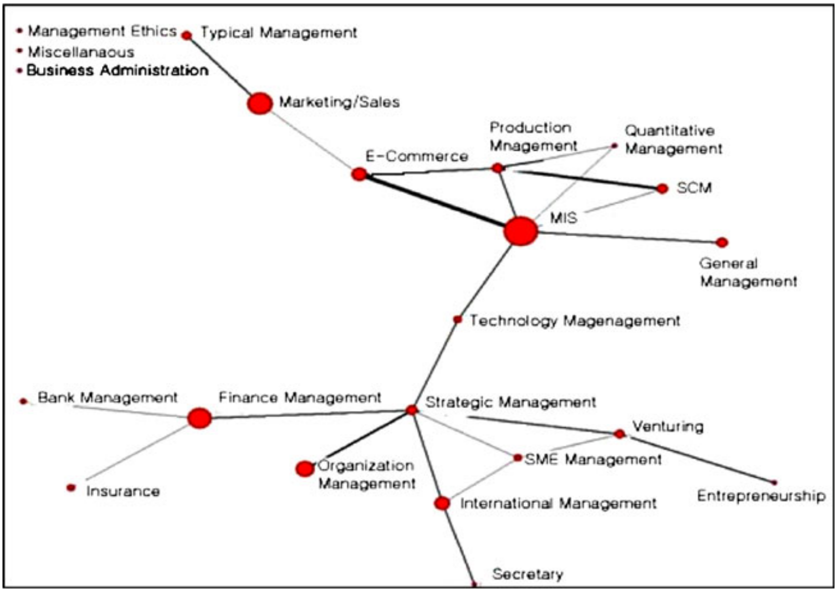

Fig. 6 R&D relation map in the 'Management' category

Table 2 List of highly ranked research areas with regard to centrality

Rank	Degree centrality		Betweenness centrality	
	Research area	Value	Research area	Value
1	Management strategy/policy	28.571	Management strategy/policy	50.476
2	MIS	28.571	MIS	41.19
3	Production management	19.048	Technology management	39.095
4	Entrepreneurship/venturing	14.286	Financing	15.714
5	International management	14.286	Human resource management	15.238

of an R&D cluster map in the 'management' research area. Clustering analysis is applied to derive groups of subcategories, resulting in 12 clusters from 67 subcategories. Figure 6 shows that a category (such as production management or SCM) in the current classification system can be subdivided into different subcategories. The result of the map will enable the revision of an existing classification and help to reflect the real relations between categories through content analysis.

Conclusions and future research

This research aims at developing a brand new knowledge map in relation to R&D in order to support R&D management by analyzing the database of research proposals. To this end, instead of using simple statistics related to research information, content analysis is conducted to extract key research subjects. Thus, a keyword-based KM is generated by constructing a similarity matrix of keyword vectors. In addition, various forms of keyword-

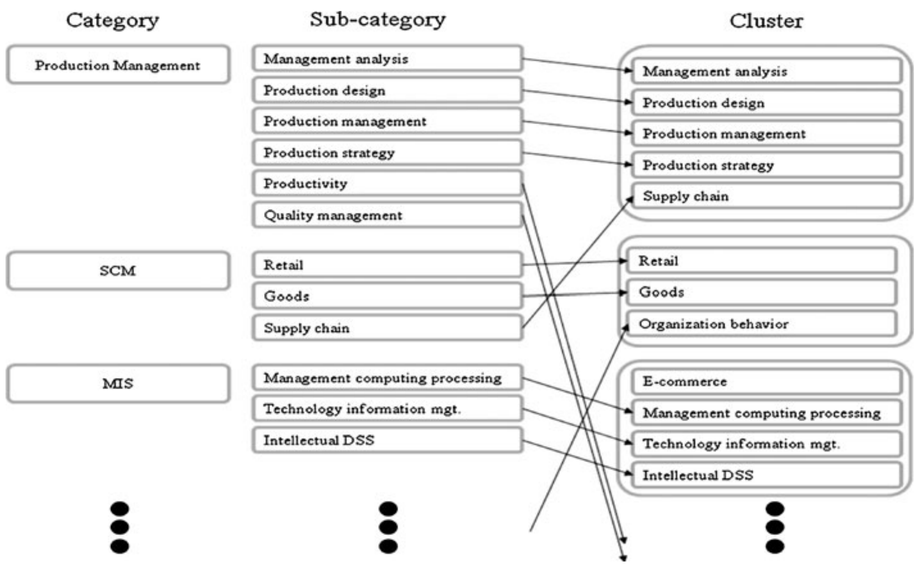

Fig. 7 R&D cluster map in the 'Management' category

based KM are suggested, e.g., maps, networks, and graphs. The types of keyword-based KM span the core R&D map, R&D trend map, R&D concentration map, R&D relation map, and R&D cluster map. These KMs can assist researchers in identifying important R&D areas, investigating relationships across various R&D areas, and understating the trends and groups of R&D subjects.

The advantages of the proposed maps are as follows. First of all, although existing bibliometric maps tend to analyze patents or academic papers in order to investigate the trends of research activities, the keyword-based knowledge maps in this paper utilize research proposals that are submitted to the NRF. Thus, the collected data to draw knowledge maps on research are appropriate to examine the national interests on the trends of promising research areas. Second, the proposed knowledge maps enable policy makers to grasp the research status of universities and countries by visualizing the patterns and trends of research activities. In particular, the maps can be used as fundamental data to allocate R&D budgets in valuable research themes and promote specialized research programs. Moreover, the knowledge maps are regarded as an effective method to balance the R&D portfolio of countries by positioning existing research programs. Thus, the characteristics of research focuses of a country can be compared to those of other countries. Third, the proposed maps to use keywords and various indices such as HI and centrality measures can present a new approach to process raw data (research proposals) and investigate the characteristics of research activities. A visual form itself focuses on providing a big picture, enabling qualitative analysis. However, since more implications can be derived from quantitative data, such indices enable researchers to conduct detailed analysis.

However, this exploratory research is subject to several limitations. First, a process of data cleansing was not clearly implemented. Although keywords were extracted exclusively and collectively, clusters of words that had similar meanings were not integrated in this research. Second, the hierarchy of keywords was not considered for generating keyword vectors. Such a hierarchical structure might be critical for correctly analyzing relations between research areas. Third, while core R&D subjects, trends, and clusters of R&D

are presented, this paper does not identify concrete and promising R&D areas that might contribute more to R&D management. Thus, future research should be conducted to overcome the presented limitations. First, a methodology for constructing the ontology of keywords needs to be suggested so that critical keywords can be extracted systematically. Second, an automated process to build the ontology, extract keywords, and draw a KM should be proposed as a practical tool. Finally, a function to monitor research trends and forecast promising research subjects needs to be developed to support R&D project selection. The results of this research can help researchers and policy makers to investigate the trends in the R&D areas of interest and allocate research budgets to promising R&D projects.

Acknowledgments This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) and funded by the Ministry of Education, Science, and Technology (Grant No. 2009-0073285).

References

Aaronson, S. (1975). The footnotes of science. *Mosaic, 6*, 22–27.

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy, 38*, 206–215.

Abramo, G., D'Angelo, C. A., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics*, *76*(2), 225–244.

Aksnes, D. W., & Taxt, R. E. (2004). Peers reviews and bibliometric indicators: A comparative study at Norwegian University. *Research Evaluation, 13*(1), 33–41.

Borner, K., Hardy, E., Herr, B., Hooloway, T., & Paley, W. B. (2007). Taxonomy visualization in support of the semi-automatic validation and optimization of organizational schemas. *Journal of Informetrics, 1*, 214–225.

Boyack, K. W., & Klavans, R. (2008). Measuring science–technology interaction using rare inventor–author names. *Journal of Informetrics, 2*, 173–182.

Browne, G., Curley, S., & Benson, P. (1997). Evoking information in probability assessment: Knowledge maps and reasoning-based directed questions. *Management Science, 43*(1), 1–14.

Calero-Medina, C., & Noyons, E. C. M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics, 2*, 272–279.

Chen, C. (1999). *Information visualization and virtual environments*. Berlin: Springer.

Chen, C., & Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer, 34*(3), 65–71.

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

Cohen, W. W., & Hirsh, H. (1998). Joints that generalize: Text classification using WHIRL. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, New York.

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning, 42*(1), 143–175.

Dixon, M. (1997). *An overview of document mining technology*. Unpublished paper.

Doyle, L. B. (1961). Semantic roadmaps for literature searchers. *Journal of the Association for Computing Machinery, 8*, 553–578.

Garfield, E. (1963). Citation indexes in sociological and historical research. *American Documentation, 14*, 289–291.

Goldstone, R. L., & Kersten, A. (2003). Concepts and categories. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive handbook of psychology* (pp. 591–621). New York: Wiley.

Griffith, B. C., Small, H., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literature, II: Toward a macro and microstructure for science. *Science Studies, 4*, 339–365.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(5–6), 907–928.

Hammer, H., Garcia-Molina, J., Cho, R., Aranha, A., & Crespo, V. (1997). Extracting semistructured information from the Web. In *Proceedings of the Workshop on Management of Semistructured Data (PODS/SIGMOD'97)*, Tucson, AZ.

Hayashi, T. (2003). Bibliometric analysis on additionality of Japanese R&D programmes. *Scientometrics, 56*(3), 301–316.

Herl, H. E., O'Neil, H. F. J., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computer in Human Behavior, 15*(3), 315–333.

Hsu, N. N., & Chang, C. C. (1999). Finite-state transducers for semi-structured text mining. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) workshop on text mining*, Stockholm, Sweden.

Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text classification using graph mining-based feature extraction. *Knowledge-Based Systems, 23*, 302–308.

Klavans, R., & Boyack, K. W. (2007). Is there a convergent structure to science? In D. Torres-salinas & H. F. Moded (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics* (pp. 437–448). Madrid: CSIC.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Leydesdorff, L., Cozzens, S., & Van Den Besselaar, P. (1994). Tracking areas of strategic importance using scientometric journal mappings. *Research Policy, 23*(2), 217–229.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*, 348–362.

Li, Y., Chung, S. M., & Holt, J. D. (2008). Text document clustering based on frequent word meaning sequences. *Data & Knowledge Engineering, 64*, 381–404.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*, 354–368.

Lourenco, A., Carreira, R., Carneiro, S., Maia, P., Glezopena, D., Fdez-Riverola, F., et al. (2009). @Note: A workbench for biomedical text mining. *Journal of Biomedical Informatics, 42*, 710–720.

Maule, R. W. (1997). Cognitive maps, AI agents and personalized virtual environment in internet learning experience. *Internet Research, 8*(4), 347–358.

Moed, H. F., De Bruin, R. E., Nederhof, A. J., & Tijssen, R. J. W. (1991). International scientific co-operation and awareness within the European community: Problems and perspectives. *Scientometrics, 21*, 291–311.

Moon, Y. H. (2005). *Next-generation information analysis*. Daejeon, Korea: KISTI.

Murray, F. (2002). Innovation as co-evolution of scientific and technological networks: Exploring tissue engineering. *Research Policy, 31*, 1389–1403.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: National Science Foundation.

Narin, F., & Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics, 36*(3), 293–310.

National Institute of Science and Technology Policy (NISTEP). (2007). *Science map 2004: Study on hot research areas (1999–2004) by bibliometric method*. Tokyo, Japan: NISTEP.

Novak, J. D. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and cooperations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Noyons, E. C. M. (1999). *Bibliometic mapping as a science policy and research management tool*. Thesis, Leiden University. Leiden: DSWO Press.

Noyons, E. C. M., Luwel, M., & Moed, H. F. (1998). Assessment of Flemish R&D in the field of information technology - A bibliometric evaluation based on publication and patent data, combined with OECD research input statistics. *Research Policy*, *27*, 287–302.

Perry, C. A., & Rice, R. E. (1998). Scholarly communication in developmental dyslexia: Influence of network structure on change in a hybrid problem area. *Journal of the American Society for Information Science, 49*(2), 151–168.

Price, D. J. D. (1965). Networks of scientific papers. *Science, 149*, 510–515.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation, 25*(4), 348–349.

Santo, M. M., Coelho, G. M., Santos, D., & Filho, L. (2006). Text mining as a valuable tool in foresight exercises: A study on nanotechnology. *Technological Forecasting & Social Change, 73*, 1013–1027.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Small, H. G. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science*, *7*, 139–166.

STEPI. (2003). *Development and application of knowledge maps for analysis of planning new technology*. Seoul, Korea: STEPI.

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

Vail, E. F. (1999). Mapping organizational knowledge. *Knowledge Management Review, 8*, 10–15.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*, 397–420.

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133–143.

Van Raan, A. F. J., & Van Leeuwen, Th. N. (2002). Assessment of the scientific basis of interdisciplinary, applied research application of bibliometric methods in nutrition and food research. *Research Policy, 31*, 611–632.

White, H. D. (2001). Author-centered bibliometircs through CAMEOs: Characterization automatically made and cited online. *Scientometrics, 51*, 607–637.

White, H. D., Buzydlowski, J., & Lin, X. (2000), Co-cited author maps as interfaces to digital libraries: Designing pathfinder networks in the humanities. In *Proceedings of the IEEE International conference on Information Visualization*, London, UK.

White, H. D., Lin, X., & Mccain, K. W. (1998). Two modes of automated domain analysis: Multidimensional scaling vs. Kohones feature mapping of information science authors. In *Proceedings of the Fifth International ISKO Conference* (pp. 57–61). Wurzberg, Germany: Ergon Verlag.

Yang, Y. (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, Dublin, Ireland.

Yang, H., & Lee, C. (2004). A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications, 27*, 645–663.

Yoon, B. U. (2008). Structuring technological information for technology roadmapping. In *Proceeding of the 7th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Databases (AIKED '08)*, Cambridge, UK.

Yoshiyuki, T., Shiho, M., Yuya, K., & Katsumori, M. (2009). Nanobiotechnology as an emerging research domain from nanotechnology: A bibliometric approach. *Scientometrics, 80*(1), 25–40.