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

Article *in* The Journal of High Technology Management Research · February 2004

DOI: 10.1016/j.hitech.2003.09.003

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of High Technology
Management Research 15 (2004) 37–50

THE JOURNAL
OF HIGH
TECHNOLOGY
MANAGEMENT
RESEARCH

A text-mining-based patent network: Analytical tool for high-technology trend

Byungun Yoon, Yongtae Park*

Department of Industrial Engineering, School of Engineering, Seoul National University, San 56-1,
Shillim-Dong, Gwanak-Gu, Seoul 151-742, South Korea

Accepted 22 September 2003

Abstract

Patent documents are an ample source of technical and commercial knowledge and, thus, patent analysis has long been considered a useful vehicle for R&D management and technoeconomic analysis. In terms of techniques for patent analysis, citation analysis has been the most frequently adopted tool. In this research, we note that citation analysis is subject to some crucial drawbacks and propose a network-based analysis, an alternative method for citation analysis. By using an illustrative data set, the overall process of developing patent network is described. Furthermore, such new indexes as technology centrality index, technology cycle index, and technology keyword clusters are suggested for in-depth quantitative analysis. Although network analysis shares some commonality with conventional citation analysis, its relative advantage is substantial. It shows the overall relationship among patents as a visual network. In addition, the proposed method provides richer information and thus enables deeper analysis since it takes more diverse keywords into account and produces more meaningful indexes. These visuals and indexes can be used in analyzing up-to-date trends of high technologies and identifying promising avenues for new product development.

Keywords: Patent network; Technoeconomic analysis; Citation analysis

1. Introduction

Patent documents are an ample source of technical and commercial knowledge in terms of technical progress, market trend, and proprietary ownership and, thus, patent analysis has long been considered as a useful vehicle for R&D management in corporate setting and technoeconomic analysis in macro

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

doi:10.1016/j.hitech.2003.09.003

context. Furthermore, patents facilitate analytical work due to their relative advantages, vis-à-vis other indexes, with respect to availability of database, scope of coverage, and richness of information (Kuznets, 1962; OECD, 1994). Recently, the strategic importance of patent analysis is more highlighted in high-technology management as the process of innovation becomes more complex, the cycle of innovation becomes shorter, and the market demand becomes more volatile.

The research spectrum and practical applicability of patent analysis is wide among a variety of technology agents—R&D managers, academicians, and policy makers. In a macro sense, patent analysis has often been employed to generate economic indicators that gauge the linkage between technology development and economic growth (Grandstrand, 1999; Grilliches, 1990; Holl, Jaffe, & Trajtenberg, 2000), estimate technological knowledge flows and their impact on productivity (Evenson & Puttnam, 1988; Scherer, 1982), or compare innovative performance in international context (Paci, Sassu, & Usai, 1997). In a micro level, patent analysis has been used to evaluate the competitiveness of firms (Narin & Noma, 1987), develop technology plans (Mogee, 1991), prioritize R&D investment (Hirschey & Richardson, 2001), or monitor technological change in firms (Archibugi & Pianta, 1996; Basberg, 1987).

In general, patent analysis utilizes bibliometric data that include such information as patent number, type of document, title, inventor, international patent classification, date of application, and so forth (Gupta & Pangannaya, 2000). Then, a number of techniques may be used to manipulate and analyze bibliometric data. Amongst others, the most frequently adopted tool is patent citation analysis (Narin, 1994). Patent citations are defined as the count of citations of a patent in subsequent patents, and citations per patent represent the relative importance of the patent. Based on this idea, patent citation analysis executes a bibliometric analysis on patent documents. In essence, the methodology is a citation-based technique in that it attempts to link patents in a patent database in the same way as science citation analysis links references in a scientific paper database (Karki, 1997). Ultimately, patent citation analysis produces such technological indexes as citations per patent, highly cited patents, nonpatent link, technical impact index, current impact index, technology cycle time, and so forth. These indexes then have been used as measures of quality of technical assets (Hirschey & Richardson, 2001), negotiation power between firms (Mowery, Oxley, & Silverman, 1998), economic value of innovative outputs in market value equation (Holl et al., 2000), or domestic or cross-border technology linkages and knowledge flows (Tijssen, 2001).

Patent citation analysis, albeit easy to understand and simple to use, is subject to some serious drawbacks. First, it is very difficult to grasp the overall relationship among all the patents because citation analysis merely indicates individual links between two particular patents. Second, related to the first problem, the scope of analysis and the richness of potential information are limited because citation analysis takes only citing–cited information into account. Third, citation analysis has no capability of considering internal relationship between patents. It takes only existence or frequency of citations into account and hence may produce superficial or even misleading indexes. Finally, citation analysis is a time-consuming task because it needs an exhaustive search.

Recognizing the shortcomings of citation analysis, the main objective of current research is to propose a network-based patent analysis, an alternative method for patent citation analysis. Although network analysis shares some commonality with citation analysis, its relative advantage is substantial. First, network analysis shows the relationship among patents as a visual network and therefore assists the analyzer in intuitively comprehending the overall structure of a patent database. Second, network analysis enriches the potential utility of patent analysis because it takes more diverse keywords into account and produces more meaningful indicators. Third, the proposed method is more economical, in terms of search time and cost, because it transforms original documents into structured data through text mining technique.

This paper is organized as follows. First, as an introductory statement, the general background of network analysis and text mining is presented. Next, the overall process of developing patent network and conducting patent analysis is described. To this end, such new indexes as technology centrality index, technology cycle index, and technology keyword clusters are suggested. Then, an exemplary case is used to exhibit the process of analysis and to assure the utility of application. Finally, implications of current research and issues of future research are discussed.

2. Theoretical background

As mentioned before, the underlying theory of patent network is due to network analysis and text mining. Therefore, the basic background of network analysis and text mining is presented briefly.

2.1. Network analysis

In general, for a set of actors, the interactive relationship among actors can be portrayed as a network (Gelsing, 1992). Network analysis, a quantitative technique derived from graph theory, facilitates the analysis of interactions (edges) between actors (nodes). Actors may be discrete individuals of any kind and relation is the collection of ties among actors in the group. Usually exhibited as a visual form, the structure of relations among actors and the location of individual actors in the network provide rich information on the behavioral, perceptual, and attitudinal aspects of individual units and the system as a whole (Knoke & Kuklinski, 1982; Marseden & Laumann, 1984).

The applicability of network analysis is wide and diverse. Some typical topics include the world political and economic system (Synder & Kick, 1979), interindustrial diffusion and adoption of innovations (Leoncini, Maggioni, & Montressor, 1996; Park & Kim, 1999), and human network in knowledge management (Cross, Borgatti, & Parker, 2001).

In the context of patent analysis, individual patents account for nodes and the relationships among patents represent edges in the network. The intrinsic connectivity between patents is not cartographical in text format, but by visualizing the locations of individual patents and linkage patterns among patents, it becomes possible to view the overall landscape on a global scale and from different perspectives.

2.2. Text mining

Data mining, also known as knowledge discovery in a database, is a recent development for accessing and extracting information in a database (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996; Piatetsky-Shapiro & Frawley, 1991). In short, data mining applies machine-learning and statistical analysis techniques for the automatic discovery of patterns in a database. Most efforts in data mining, however, have been made to extract information from a structured database and the utility of data mining is yet limited in handling huge amounts of unstructured textual documents.

As a remedy, text mining is a rather new technique that has been proposed to perform knowledge discovery from collections of unstructured text. In short, text mining puts a set of labels on each document and discovery operations are performed on the labels. The usual practice is to put labels to words in the document. Then, the document in text format can be featured by keywords that are extracted through text mining algorithm. Recently, text mining has attracted increasing interest and has been actively applied in knowledge management (Feldman et al., 1998). A more detailed and comprehensive review of text mining can be found in Kostoff, Toothman, Eberhart, and Humenik (2001).

In relation to patent analysis, text mining is used as a data-processing and information-extracting tool. Since the original patent documents are expressed in text (natural language) format, it is necessary to transform raw data into structured data. Then, the process of keyword extraction is applied to identify keywords and to measure similarity between patents.

3. Data source

The data source of current research is about wavelength division multiplexing (WDM)-related patents. This relatively new technology enhances data processing rate significantly by carrying data over a single fiber by the use of multiple wavelengths, each carrying a separate channel. Patent documents are extracted from the U.S. Patent and Trademark Office (USPTO: www.uspto.gov) database.

In all, 119 patents were collected with the reference period from 1987 to 2001. The subject set contains patents from U.S. Patent No. 4,677,618 to U.S. Patent No. 6,333,798. However, because the real patent number is too long to be displayed on the network, serial numbers from 1 to 119, sorted by the application date, are labeled on each patent. The reason why WDM-related patents are selected is as follows. First, WDM is an emerging technology that is developed to meet the explosive demands for bandwidth in optical network. Therefore, the patent network contains dense links and active flows of knowledge among patents. Second, the size of data set is suitable, neither large nor small. If the size is too large, it becomes an intractable task to draw patent network due to space problem. If the size is too small, on the other hand, the network provides little information. Nevertheless, the WDM-related data set is used only for the purpose of illustration and the generic methodology of network analysis is applicable to any patent set.

4. Process of network-based patent analysis

The overall process of conducting network-based patent analysis goes through several steps. First of all, data collection and data preprocessing are the preliminary step. The patent area of interest is selected and related patent documents are collected in electronic text format. Second, raw patent documents are transformed into structured data. Since the original documents are expressed in natural language format, they must be transformed into structured data in order to be analyzed and utilized. Text mining that extracts keywords from patent document is used to this end. Third, patent network is generated with nodes (patents) and links (relation among patents). Intuitive but comprehensive analysis may be possible by investigating the visual pattern of network. Finally, based on some quantitative indexes, in-depth patent analysis is carried out to obtain quantitative information for decision making.

Fig. 1 depicts the overall process of generating patent network and executing patent analysis. More detailed explanation for each step is provided below. We presume that readers who are unfamiliar with various methods used here may find it difficult to understand technical details. However, under space

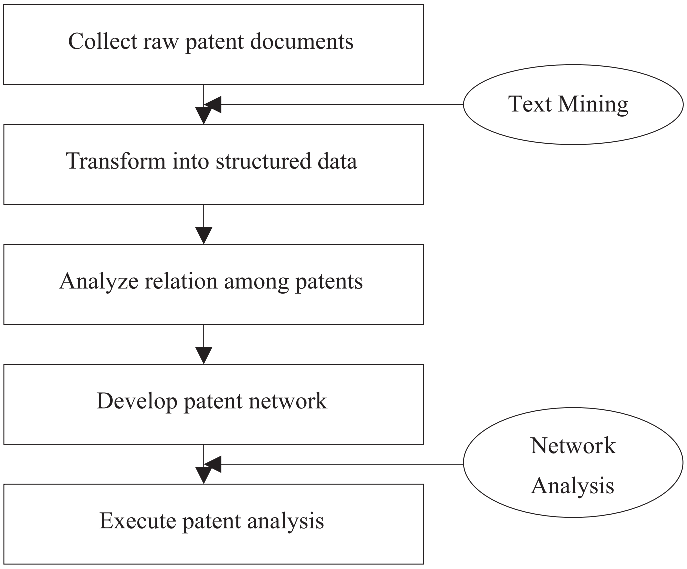
Collect raw patent documents
Text Mining
Transform into structured data
Analyze relation among patents
Develop patent network
Network Analysis
Execute patent analysis

Fig. 1. Overall process of network-based patent analysis.

constraint, the readability may be sacrificed to some extent since the main purpose of this paper is not to expatiate on a particular methodology but to propose an overall framework of development process.

4.1. Data preprocessing: keyword vector

As pointed out before, raw documents need to be preprocessed because they are unstructured in format. In the current research, text mining is executed to transform raw documents into keyword vector data and incidence matrix. Specifically, the process of keyword extraction consists of the following four stages. First, supplementary words are eliminated because they carry no semantic meaning. Second, word stems are identified and these stems are separated from their accompanying prefixes and suffixes. This procedure facilitates analysis of stems that yields a clearer picture in terms of frequency and relationship in the text. Third, keywords in the document are identified. To this end, statistical analysis, frequency analysis in particular, is conducted first to set individual weights of words and weights of respective relations of each other. Then, word stems with high frequency are determined as keywords. Finally, keyword vector is constructed. If a specific keyword is included in a patent document, then the corresponding keyword vector field is filled with frequency of occurrence.

Fig. 2 shows examples of transformed vectors for a selective set of patents: Patent 1 (USTPO No. 4,677,618), Patent 2 (USTPO No.4,747,649), and Patent 119 (USTPO No. 6,333,798). For instance, in the document of Patent 1, the word “optical” occurs 63 times, “wavelength” 19 times, and so on.

4.2. Construction of incidence matrix

Based on the keyword vector, the incidence matrix is constructed as a prerequisite for generating network. To construct the incidence matrix, the relationship between patents should be quantified in terms of either distance or similarity. Among various association indexes, the common Euclidian

Patent 1	:	(63, 19, 35, 13, 61, 42, 2, 40, 1, ..., 15)
Patent 2	:	(1, 8, 51, 8, 0, 0, 0, 0, , ..., 0)
⋮		⋮
Patent 119	:	(578, 21, 62, 61, 0, 96, 45, 5, 0, ..., 39)

Fig. 2. Examples of keyword vector.

distance index is used in this research (Johnson & Wichern, 1988). If keyword vectors of n_i and n_j are defined as (n_{i1}, n_{i2}, . . ., n_{ik}) and (n_{j1}, n_{j2}, . . ., n_{jk}), respectively, the association value between two vectors (nodes) is computed as follows:

$$\text{Association value} = \sqrt{\frac{(n_{i1} - n_{j1})^2 + (n_{i2} - n_{j2})^2 + \cdots + (n_{ik} - n_{jk})^2}{k}}$$

Although the association values assume real numbers from 0.0 to 1.0, the incidence matrix contains binary values where entry I_{ij} equals 1 if node i is strongly connected with node j but equals 0 if node i is not or loosely connected with node j. The degree of connectivity, whether strong or weak, is decided based on the cutoff value that the analyzer is supposed to determine. That is, the connectivity between node i and node j is considered strong and set to 1 in the incidence matrix if the association value between two nodes is larger than cutoff value. Otherwise, the connectivity is considered weak and set to 0. The determination of cutoff value is in nature a subjective, trial-and-error task. The network becomes denser as the cutoff value becomes lower, whereas it becomes sparser as the cutoff value becomes higher. At some intermediate level, the analyzer has to select a reasonable value so that the structure of the network becomes clearly visible. However, in general, it is advisable to apply multiple values for a sensitivity analysis.

4.3. Generation of patent network

By applying the incidence matrix as input, one can readily use some networking software packages to generate patent networks. In this research, we used UCINET 5 and Krackplot 3.0, software for network and graph.

For the purpose of illustration, the cutoff value of 0.16, 0.17, and 0.18 are selected in the current research. Figs. 3, 4, and 5 exhibit the patent networks of WDM-related patents for each cutoff value. Note that the structure of network is very sensitive to cutoff value. In this particular case, the cutoff value of 0.17 seems to generate the most visible and meaningful network.

A well-constructed visual display of network often conveys an intuitive knowledge on the structure of a system (Knoke & Kuklinski, 1982). Likewise, patent network provides some valuable insights into the holistic nature of a subject set. For instance, Fig. 3 provides the overall view of network in terms of connectivity. First, the network divides all the patents into two sets, interconnected set and isolated set. This information facilitates separate, in-depth analysis on two different sets. Furthermore, the visual diagram indicates the distinctive role and relative importance of individual patents in the family. To illustrate, for the interconnected set, it is visible that Patents 58, 110, 107, and 115 are

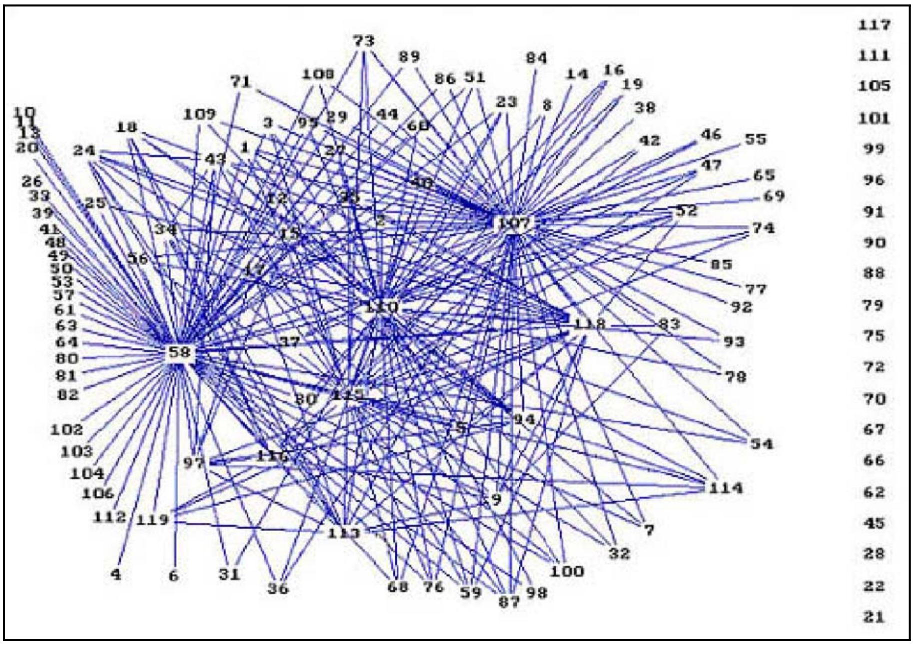

Fig. 3. Patent network of WDM-related patents: cutoff value = 0.16.

central points of knowledge flow. Patents 43, 97, 113, 114, and 118 seem to serve as intermediary points. Other patents may be named as peripheral points. As recommended, it may be desirable to obtain a number of different snapshots by changing the level of cutoff value.

4.4. Quantitative analysis of patent network

The patent network provides intuitive knowledge on the overall structure of patents for a given set of patents. To carry out more detailed analysis and to obtain richer information, however, quantitative indexes need to be operationally defined and actually gauged. Although various indexes can be developed to this end, as summarized in Table 1, we propose three major dimensions of analysis and operationally defined related index for each dimension. First, the "importance" dimension of a subject patent measures the density of linkage with other patents. Second, the "newness" dimension of subject technology indicates the average age of linked patents. Third, the "similarity" dimension combines subject patents with other similar patents to form a patent package. In correspondence, such new indexes as technology centrality index, technology cycle index, and technology keyword clusters are suggested for each dimension.

4.4.1. Technology centrality index

In patent citation analysis, a number of citation indexes are developed and used for quantitative analysis. Amongst others, the current impact index is a typical measure to indicate the frequency of citation in other patents as the foundation for invention. This measure is calculated based on how frequently the subject patent has been cited in the past 5 years' other patents and is used to gauge the impact of the subject patent.

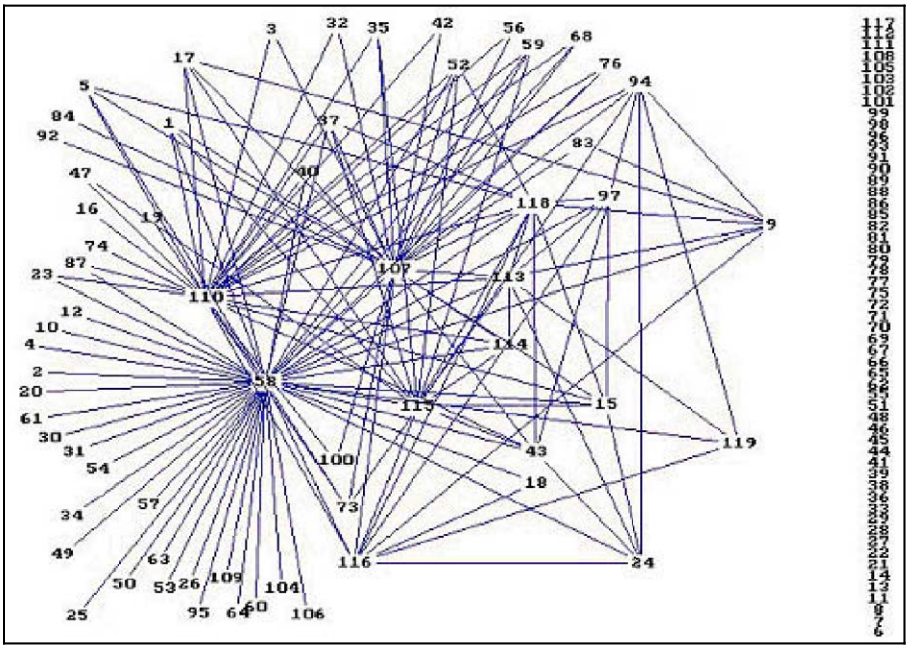

Fig. 4. Patent network of WDM-related patents: cutoff value = 0.17.

In network-based analysis, instead, technology centrality index is proposed. The centrality index, equivalent to actor degree centrality in network analysis, is defined as:

$$C_{\mathrm{D}}(n_i) = \frac{d(n_i)}{g-1}$$

where $d(n_i)$ is the number of lines that are incident with patent i and g is the total number of patents.

The centrality index in patent network is interpreted as the ratio of the number of tied links to all $g-1$ other patents. Therefore, the higher the centrality index, the greater the impact on other patents. Conceptually, this index is similar to current impact index of citation analysis, but practically it is more informative and reliable. In citation analysis, the existence of relationship between two patents is determined by the existence of citing–cited records and the relative importance of a patent is measured by the frequency of citation by other patents. The frequency, however, would not reveal the internal, structural relationship between two patents and thus the value may be superficial or even spurious. In network analysis, the intrinsic characteristics of a patent are represented by a keyword vector and the relationship between two patents is judged by the similarity between two vectors. Therefore, the link between two patents reflects not only the value of frequency (quantity) but also the value of association (quality).

The centrality indexes of some selective patents are presented in Table 2. For the purpose of illustration, those patents with index value higher than 0.1 are listed. As intuitively identified in the visual network, Patent 58 shows the highest value and thus turns out to be the most influential patent. In detail, this patent, U.S. Patent No. 6,038,357, is entitled PDM–WDM for fiber optic communication

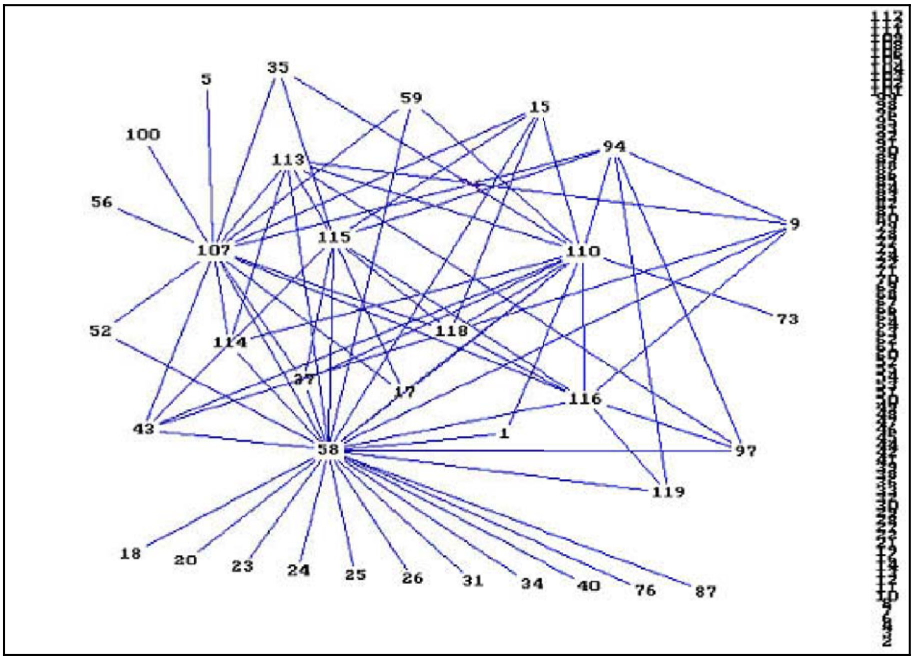

Fig. 5. Patent network of WDM-related patents: cutoff value = 0.18.

networks. This invention is related to the field of fiber optic systems and networks in which signals are directed from an optical source to a receiver by the state of polarization and the wavelength of the optical signal. Since this patent combines polarization-division multiplexing (PDM) and wavelength-division multiplexing (WDM) to increase capacity, it naturally exhibits very active inflow and outflow of knowledge across other patents. As anticipated, Patents 110, 107, 115, and 118 also show relatively high values.

To compare network analysis to conventional citation analysis, we computed current impact indexes of citation analysis, to find out that the result is quite different from each other. In citation analysis, Patents 24, 11, 26, 18, and 8 show the highest impact index values. Note that the serial numbers of patents are sorted by the application date and, thus, these patents are older ones, as compared to those of the centrality index. This result seems to be quite natural since citation analysis is age dependent in that old patents have more chances to be cited and therefore have more possibility to exhibit higher impact value. In other words, no matter how influential they are, rather new patents tend to show lower impact

Table 1
Dimension of analysis and related indexes

Dimension	Index	Objective and definition
Degree of importance	Technology centrality index	Measure the relative importance of subject patent by gauging density of linkage with other patents.
Degree of newness	Technology cycle time	Measure the status in life cycle of subject technology by gauging average age of linked patents.
Degree of similarity	Technology keyword cluster	Group similar patents together to form patent package.

Table 2
Technology centrality indexes of some important patents

Patent number (real patent number)	Centrality index value
58 (U.S. Patent No. 6,038,357)	0.42
110 (U.S. Patent No. 6,307,986)	0.25
107 (U.S. Patent No. 6,288,812)	0.22
115 (U.S. Patent No. 6,321,004)	0.14
118 (U.S. Patent No. 6,327,076)	0.11

values simply because they have less chance to be cited. In that regard, we believe that the result of citation analysis may lack the capability to reflect the "real" influence of patents and, as a remedy, advocate the use of the centrality index of network analysis.

In practice, the centrality index can be used to identify the list of influential patents. Once identified as influential, detailed information on the subject patent needs to be retrieved, and technical and strategic implication needs to be drawn. Generally, patents may be categorized into several groups depending on the index value, and the degree of management should be differentiated across categories.

4.4.2. Technology cycle index

In citation analysis, cycle time is defined as the median age, in years, of the earlier patents that are cited in the subject patent. In general, the age of patents referenced by another patent reflects the chronological history of the prior art on which the citing patent is built. Thus, the index may indicate the speed of technical advancement and/or the newness of subject patent. That is, if the cycle time of the subject patent is shorter than that of another, it means that the subject patent is built on more recent technologies or is facing more rapid technical changes. The cycle varies from 4 to 5 years for the rapidly changing electronics areas to more than 15 years for some of the gradually moving mechanical areas (Karki, 1997).

In this research, we propose the technology cycle index as an alternative of cycle time of citation analysis. The cycle index of specific technology is defined as the median of age gaps between subject patent and other connected patents. Although seemingly identical to cycle time of citation analysis, the proposed index is differentiated due to the following two aspects. First, again, the new index represents the median age of "connected" patents, not of "cited" patents. Therefore, it will indicate the trend of technological advance more accurately. Second, the new index takes age gaps of both earlier citations and later citations into consideration. Note that the cycle time of citation analysis represents the median age of the earlier patents that are cited in the subject patent. However, technology cycle index here indicates the median age of both the earlier patents that are cited in the subject patent and the later patents, if any, that have cited the subject patent. By considering the whole set of related patents, it is expected that the index may reflect the technical trend more accurately and comprehensively.

Table 3 contrasts cycle indexes of two groups, shorter cycle group and longer cycle group. In the table, those patents with cycle time longer than 3 years are classified as long-cycle-time group, whereas those with cycle time shorten than 1 year are assigned to short-cycle-time group. It is noteworthy that patents with lower number (earlier in application date) tend to exhibit longer cycle while those with higher number (recent in application date) show shorter cycle time. This phenomenon may indicate the general trend that technology cycle becomes shorter over time. However, although not presented in this paper, wide variations are also found across patent families irrespective of application date.

Table 3
Technology cycle indexes of some selective patents

Short technology cycle time group	Cycle time (year)	Long technology cycle time group	Cycle time (year)
Patent number (real number)		Patent number (real number)	
114 (6,320,684)	0.2	1 (4,677,618)	13.75
119 (6,333,798)	0.25	5 (5,101,290)	8.75
59 (6,043,914)	0.75	84 (6,195,186)	4
73 (6,141,370)	0.75	17 (5,668,652)	3.8
97 (6,243,177)	1	15 (5,625,478)	3.43

As an attempt to compare the current cycle index to conventional cycle index, we computed values of both indexes. As a whole, interestingly, there appears a substantial difference in case of shorter cycle patents while there is found a negligible gap in case of longer cycle patents. To illustrate a few, for Patent 1 in the longer cycle group, the current index value is 13.75, whereas that of the conventional index is 13.0. However, for Patent 119 in the shorter cycle group, the former is 0.25 while the latter is 1.4. Obviously, the difference in value is due to the difference in operational definition of index. Again, we believe that the age of "connected" patents reflects the trend of technological advance more accurately than that of "cited" patents.

In high-technology management, the changing trend of technical advancement should be carefully monitored. The cycle index provides useful information to this end. The utility of the cycle index may be extended further. It enables comparative analyses between technology groups, age groups, or industrial sectors.

4.4.3. Technology keyword clusters

In citation analysis, co-citation clusters are often drawn in order to make patent groups. By definition, two patents are co-cited if they are both referenced by another patent. The number of times two documents are co-cited by other referencing documents is called the strength of co-citation link. The stronger the link is, the closer the interrelationship is between the two. This information provides a strategic possibility to group co-cited patents as a cluster and license them together as a package.

Likewise, we propose the technology keyword clusters in this research. The keyword cluster technique defines the similarity of two patents as the similarity in terms of keyword vectors, instead of frequency of co-citations. By doing so, it is expected that patents are assigned to more appropriate partitions and thus more intrinsically homogeneous clusters are produced. As explained before, co-citation analysis deals with the patents as a whole and considers only the frequency of co-citation. Thus, the result of grouping may be superficial or even spurious since the co-citation statistics itself would not reveal the internal, structural relationship between patents.

The current research applies the K-means clustering algorithm that has a predefined number of clusters (Johnson & Wichern, 1988). Among various possible numbers of clusters (values of K), the best number tends to show the smallest intracluster distance and largest intercluster distance. If so, the elements of each cluster have a high degree of similarity and at the same time each cluster is heterogeneously discriminated from other clusters. The former is defined as the average distance between individual elements and cluster center, whereas the latter is defined as the average distance between cluster centers. Table 4 provides the comparison between some selective numbers of clusters. Note that the case of seven clusters shows the lowest ratio to generate the best clustering outcome.

Table 4
Comparison of technology keyword clusters

Number of clusters	Intracluster distances (1)	Intercluster distances (2)	(1)/(2) ratio
3	2.23	1.40	1.59
4	2.18	1.71	1.27
5	2.07	1.53	1.35
6	2.03	1.50	1.35
7	2.01	1.67	1.20
8	1.97	1.59	1.23

Consequently, as presented in Table 5, all patents are assigned to seven clusters. Note that, within each cluster, affiliated patents are similar with respect to keyword structure, not in terms of frequency of co-citations. The comparative advantage of keyword-based grouping over citation-based grouping is clear. The citation information merely shows that two patents are “referred” or “linked” to each other. On the contrary, keyword information examines the internal relationship between patents and therefore reveals whether two patents are “similar” or “homogeneous” in terms of technical contents.

The practical applicability of keyword clustering to high-tech management is as follows. First, it renders potential chance to group similar patents as a technology package. Second, the keyword clusters enable users to monitor the overall portfolio of patents. Third, by examining the detailed relationship among patents in each cluster, one may be able to identify idiosyncratic characteristics of clusters.

4.5. Conclusions and future research

It is evident that patent documents are a valuable reservoir of technical and commercial knowledge. The potential usefulness of patent data becomes more highlighted as the process of innovation becomes more complex, the cycle of innovation becomes shorter, and the market demand becomes more volatile. Simultaneously, however, more sophisticated and comprehensive methods are required to extract latent information from the raw database. To this end, we proposed an exploratory process of generating a patent network and conducting the ensuing quantitative analysis. In terms of methodology, the process integrates text mining and network analysis.

The main objective and substantial contribution of current research is to propose a network-based patent analysis. The salient feature of network analysis is twofold. First, it is equipped with visual

Table 5
Patent grouping: case of seven clusters

Cluster	Affiliated patents
1	2, 11, 33, 36, 85, 103
2	22, 53, 64, 67, 72, 98, 108
3	6, 28, 41, 60, 63, 88, 96
4	4, 13, 27, 45, 48, 101, 111
5	10, 12, 21, 39, 66, 77, 79
6	7, 29, 50, 51, 55, 82, 91
7	Others

display. Second, it generates quantitative values of patents in terms of degree of importance, degree of newness, and degree of similarity. The results of network analysis can be used, both strategically and operationally, in high-technology management. The coverage of application is wide, ranging from new idea generation to ex post facto auditing. First, the visualized display of the network enables the analyzer to easily understand the global structure of the patent set in that it shows both the overall relations among patents and the respective positions of individual patents in the network. Second, it assists users in determining the relative importance of individual patents. The analysis generates a selective set of influential patents that deserve more intensive control. Third, it facilitates analysis of up-to-date trend of high technologies and identification of promising avenues for new product development. Fourth, it helps users to manage a patent portfolio or to package patents. This activity clearly enhances the strategic capability of patent management.

Although the proposed method is more comprehensive and flexible, especially as compared to conventional citation analysis, it is still subject to some limitations. First, it may be difficult to generate a patent network if the size of patent documents is too voluminous. Second, the patent network may be ambiguous or meaningless if the structural relationship among patents is unclear.

Nevertheless, the relative advantage of network analysis is substantial and the network-based approach can be used to overcome the drawbacks of conventional citation analysis. Furthermore, the validity and utility of the proposed method can be extended and/or elaborated far beyond the scope of current research. Some promising themes of further research include the following. First, a comparative analysis among diverse patent sets is suggested. For instance, international comparison or interindustrial comparison is necessary and welcome. Second, a time-series analysis is recommended to investigate the dynamic pattern. Third, the development of other quantitative indexes is required to expand and diversify the scope of analysis.

Acknowledgements

This research was funded by the National Research Lab (NRL) program of the Ministry of Science and Technology of Korea.

References

Archibugi, D., & Pianta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 146–451.

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

Cross, R., Borgatti, S., & Parker, A. (2001). Beyond answers: Dimensions of the advice network. *Social Networks*, *23*, 215–235.

Evenson, R., & Puttnam, J. (1988). *The Yale–Canada patent flow concordance*. New Haven, CT: Economic Growth Center, Yale University.

Fayyad, U., Piatetsky-Shapiro, P., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI Press.

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge management: A text mining approach. *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM98), Basel, Switzerland, October 1998* (pp. 9(1)–9(10)). Zurich: Swiss Life Information Systems Research.

Gelsing, L. (1992). Innovation and the development of industrial networks. In B. -A. Lundvall (Ed.), *National systems of innovation—towards a theory of innovation and interactive learning*. London: Printer.

Grandstrand, O. (1999). *The economics and management of intellectual property: Toward intellectual capitalism*. UK: Edward Elgar.

Grilliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, *28*, 1661–1707.

Gupta, V., & Pangannaya, N. (2000). Carbon nanotubes: Bibliometric analysis of patents. *World Patent Information*, *22*, 185–189.

Hirschey, M., & Richardson, V. (2001). Valuation effects of patent quality: A comparison for Japanese and US firms. *Pacific-Basin Finance Journal*, *9*, 65–82.

Holl, B., Jaffe A., Trajtenberg, M. (2000). *Market value and patent citations: A first look.* NBER Working Paper Series, Cambridge, MA.

Johnson, R., & Wichern, D. (1988). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

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

Knoke, D., & Kuklinski, J. (1982). *Network analysis*. London: Sage.

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.

Kuznets, S. (1962). Innovative activity: Problems of definition and measurement. In R. Nelson (Ed.), *The rate and direction of inventive activity*. Princeton, NJ: Princeton University Press.

Leoncini, R., Maggioni, M., & Montressor, S. (1996). Intersectorial innovation flows and national technological system network analysis for comparing Italy and Germany. *Research Policy*, *25*, 415–430.

Marseden, P., & Laumann, E. (1984). Mathematical ideas in social structural analysis. *Journal of Mathematical Sociology*, *10*, 271–294.

Mogee, M. (1991). Using patent data for technology analysis and planning. *Research-Technology Management*, *34*, 43–49.

Mowery, D., Oxley, D., & Silverman, B. (1998). Technological overlap and interfirm cooperation: Implications for the resource-based view of the firm. *Research Policy*, *27*, 507–523.

Narin, F. (1994). Patent bibliometrics. *Scientometics*, *30*(1), 147–155.

Narin, F., & Noma, E. (1987). Patents as indicators of corporate technological strength. *Research Policy*, *16*, 143–155.

OECD (1994). *The measurement of scientific and technological activities: Using patents as science and technology indicators—Patent manual.* Paris.

Paci, R., Sassu, A., & Usai, S. (1997). International patenting and national technological specialization. *Technovation*, *17*(1), 25–38.

Park, Y., & Kim, M. (1999). A taxonomy of industries based on knowledge flow structure. *Technology Analysis and Strategic Management*, *11*(4), 541–549.

Piatetsky-Shapiro, G., & Frawley, W. (1991). *Knowledge discovery in database*. Menlo Park, CA: AAAI Press.

Scherer, F. (1982). Inter-industry technology flows in the United States. *Research Policy*, *11*, 227–245.

Synder, D., & Kick, E. (1979). Structural position in the world system and economic growth 1955–1970: A multiple network analysis of transnational interactions. *American Journal of Sociology*, *84*, 1096–1126.

Tijssen, R. (2001). Global and domestic utilization of industrial relevant science: Patent citation analysis of science–technology interactions and knowledge flows. *Research Policy, 30*, 35–54 (Available: www.uspto.gov).