Available online at www.sciencedirect.com
ScienceDirect
Technovation 28 (2008) 758–775
technovation
www.elsevier.com/locate/technovation

Detecting emerging research fronts based on topological measures in citation networks of scientific publications

Naoki Shibata, Yuya Kajikawa*, Yoshiyuki Takeda, Katsumori Matsushima

Institute of Engineering Innovation, School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8656, Japan

Abstract

In this paper, we performed a comparative study in two research domains in order to develop a method of detecting emerging knowledge domains. The selected domains are research on gallium nitride (GaN) and research on complex networks, which represent recent examples of innovative research. We divided citation networks into clusters using the topological clustering method, tracked the positions of papers in each cluster, and visualized citation networks with characteristic terms for each cluster. Analyzing the clustering results with the average age and parent–children relationship of each cluster may be helpful in detecting emergence. In addition, topological measures, within-cluster degree z and participation coefficient P, succeeded in determining whether there are emerging knowledge clusters. There were at least two types of development of knowledge domains. One is incremental innovation as in GaN and the other is branching innovation as in complex networks. In the domains where incremental innovation occurs, papers changed their position to large z and large P. On the other hand, in the case of branching innovation, they moved to a position with large z and small P, because there is a new emerging cluster, and active research centers shift rapidly. Our results showed that topological measures are beneficial in detecting branching innovation in the citation network of scientific publications.

Keywords: R&D management; Research front; Bibliometrics; Citation network; Topological clustering

1. Introduction

Scientific activities are playing an increasingly important role not only in solving social problems but also as seeds of industrial innovations. Previous studies have focused on establishing the relationship between the investments and the outcomes of scientific research (e.g., Mansfield, 1972; Rosenberg, 1974; Sveikauskas, 1981; Adams, 1990; Narin and Hamilton, 1996). The importance of R&D is widely recognized as essential in promoting technological innovations, especially in science-oriented disciplines such as chemicals, pharmaceutics, and electronics (Tijssen, 2002), as well as in nutrition and food research (van Raan and van Leeuwen, 2002). There is additional empirical support for a link between scientific research and technological innovation. Jaffe and Trajtenberg showed a positive relationship between university research expenditure and local patenting rates (Jaffe, 1989; Jaffe and Trajtenberg, 1996). It is often observed that, in the innovation processes, scientists create the seeds of innovation, and then companies take up these seeds, develop technologies, and industrialize as seen in the development of the photocatalyst in Japan (Tryk et al., 2000; Hashimoto et al., 2005). Technological inventions are developed by coupling independent pieces of scientific outputs (Fleming and Sorenson, 2004). Although such a linear model is often criticized (Williams and Edge, 1996; Niosi, 1999), it still seems to form a remarkable route of technological innovations.

While the importance of scientific activities in industrial innovations has been established, it is still a controversial topic concerning how each researcher and engineer should obtain, absorb, and utilize scientific knowledge for their competitive advantage. Massini et al. (2005) discussed the difference between pioneers (innovators) and adopters (imitators). For innovators and early adopters, it is essential to detect emerging research fields promptly before other competitors enter the research domain. Currently,

*Corresponding author. Tel./fax: +81 3 5841 7672.
E-mail address: kaji@biz-model.t.u-tokyo.ac.jp (Y. Kajikawa).

0166-4972/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.technovation.2008.03.009

the main output of scientific activities still lies in a number of journal papers, and scientific publications play an important role as the primary "raw material" building scientific knowledge to accelerate technological innovation. In fact, Sorenson and Fleming observed that patents that refer to scientific materials receive more citations (Sorenson and Fleming, 2004; Fleming and Sorenson, 2004). This partially supports the hypothesis that scientific publications play an important role in accelerating technological innovation. Scientific publications constitute the generally accepted, although not always perfect, major output of the scientific activity stimulating technological innovations.

In today's increasingly global and knowledge-based economy, competitiveness and growth depend on the ability of an economy to meet fast-changing market needs quickly and efficiently through the application of new science and technology. The capacity to assimilate and apply new knowledge relies on scientific innovativeness. Therefore, for both R&D managers in companies or research institutions and policy makers, noticing emerging research domains among numerous academic papers has become a significant task. However, such a task becomes highly laborious and difficult as each research domain becomes specialized and segmented. Davidson et al. (1998) consider this situation as follows: "For most of history, mankind has suffered from a shortage of information. Now, in just the infancy of the electronic age, we have begun to suffer from information excess." There are two approaches to detecting emerging research domains and the topics discussed there (Kostoff and Schaller, 2001). One straightforward manner is the expert-based approach, which utilizes the explicit knowledge of domain experts. However, it is often time-consuming and is also subjective in the current information-flooded era. Another is the computer-based approach, which is compatible with the scale of information, and it is therefore expected to complement the expert-based approach. There is a commensurate increase in the need for scientific and technical intelligence to discover emerging research domains and the topics discussed there, even for unfamiliar domains (van Raan, 1996; Kostoff et al., 1997, 2001; Losiewicz et al., 2000; Boyack and Böner, 2003; Porter, 2005; Buter et al., 2006).

One promising approach to detect emerging research domains is to analyze the citation network of scientific publications. In his classical paper, de Solla Price (1965) originally introduced the concept of a research front, research domains under developing where papers cite each other densely. According to Price, there seems to be a tendency for scientists to cite the most recently published articles. The research front builds on recent work, and the network there becomes very tight. In a given field, a research front refers to the body of articles that scientists actively cite. Researchers have studied quantitative methods that can be used to identify and track the research front as it evolves over time. Small and Griffith (1974) represented currently activated scientific specialties as clusters of co-cited articles. Co-citation strengths between pairs of documents are computed and the documents subsequently clustered to identify the research domain. Braam et al. (1991) investigated the topics discussed in the co-cited clusters by analyzing the frequency of indexing terms and classification codes occurring in these publications. Peters and van Raan (1993a) evaluated the usefulness of co-word technique by questionnaire, and improved the clustering method for co-word map (Peters and van Raan, 1993b). Citation and publication counts are used to evaluate the significance of patents (Albert et al., 1991), scholar (Mayer et al., 2004), journal (Leydesdorff et al., 1994), emerging research domain, and nation (Zhou and Leydesdorff, 2006).

The temporal patterns of co-cited clusters are usually tracked to detect emerging fields with a variety of visualization techniques. The multidimensional scaling (MDS) plot on a two-dimensional (2-D) plane is a typical example of such visualizations (Small, 1977). However, spatial configurations in MDS do not show links explicitly. There are number of efforts to improve the efficiency of visualization such as a self-organizing map (SOM) (Skupin, 2004) and a pathfinder network (PFNET) (Chen, 1999, 2004). White et al. (2004) compared these two visualization techniques and noted that while PFNETs seem to be directive about relationships, SOMs are merely suggestive. Morris et al. (2003) used a timeline visualization of the hierarchical structure produced by clustering. The animated representation of a citation network also helps us to focus on significant movements in research fronts and emerging research fields in a broad context (Boyack et al., 2002; Chen et al., 2002). However, the detection of emerging research domains by visualization requires implicit judgment by the users. It is desirable to develop statistical measures to detect emerging fields without the help of visualizations. The aim of this paper is to develop a computational tool to detect research fronts by using the topological measures (measures representing the topological role or position of a paper in citation network) of citation networks in addition to visualization.

We performed a comparative study in two research domains. One is a study on gallium nitride (GaN), which is widely recognized as a recent prominent innovation in the fields of applied physics and material science. The other is on complex networks (CN) analysis, which is also recently recognized as pioneering a new research field. We divide the papers in each research domain into clusters, track the positions of the papers in each cluster, and visualize citation networks with characteristic terms for each cluster. Topological measures are introduced to detect emerging domains without the help of visualizations. There are distinct differences between the topological measures in the two cases. By considering the difference, we discuss the possibility of detecting emerging knowledge domains using topological measures. In the next section, we give a brief historical overview of the two research fields. In Section 3, we explain our methods of this study. In Section 4, the

results are shown. Finally, in Section 5, we discuss how our method contributes to detecting emerging knowledge domains.

2. Overview of research domains

In this paper, we performed a comparative study in the following two research domains: GaN and CN. We take these cases because they are typical examples of recent remarkable innovations having somewhat different characteristics. As explained later, research in GaN has incrementally developed in the field of applied physics (Fig. 1(a)). However, branching innovation occurs in CN (Fig. 1(b)). In the following, we briefly describe the historical background of these domains. Therefore, a comparative study on these topics might bring us fruitful outcome to discuss the effectiveness and usefulness of our method.

2.1. Gallium nitride (GaN)

During the last decade, nitride semiconductors, especially GaN, have experienced an exciting development in the field of materials science and applied physics. Within a very short period after the middle of the 1990s, researchers realized the applications of GaN as blue and green light-emitting diodes (LEDs) and ultra-violet (UV) and blue laser diodes (LDs). These products are now commercially available. Innovation in this research field motivates researchers to engage in and open huge new markets for manufacturers and customers.

However, blue luminescent devices had not been realized in the middle of the 1990s. Until 1993, the only blue light-emitting devices commercially available were based on silicon carbide (SiC), which has an indirect band gap, and is thus not capable of sufficient brightness. GaN was also a candidate for blue luminescent devices due to its wide band gap. For GaN researchers, it was clear that the most serious problem in GaN was poor film quality, which was caused by large lattice mismatch between GaN and the underlying sapphire substrate. The difficulty in p-type doping, which is necessary to work a thin film of GaN as LED, was another crucial problem. However, the latter problem was considered to be connected to the former problem (Akasaki, 1998). These shortcomings were overcome by Akasaki and co-workers in the late 1980s. They deposited a buffer layer on the sapphire substrate prior to GaN deposition (Amano et al., 1986; Akasaki et al., 1989). The introduction of a buffer layer greatly improved the film quality. Then, they also succeeded in p-type doping by low-energy electron irradiation (Amano et al., 1989), which can overcome the large acceptor activation energy of GaN. Later, the doping was improved by annealing (Nakamura et al., 1992). Owing to these breakthroughs, the first commercial LEDs finally appeared at the end of 1993 through Shuji Nakamura working for Nichia Company in

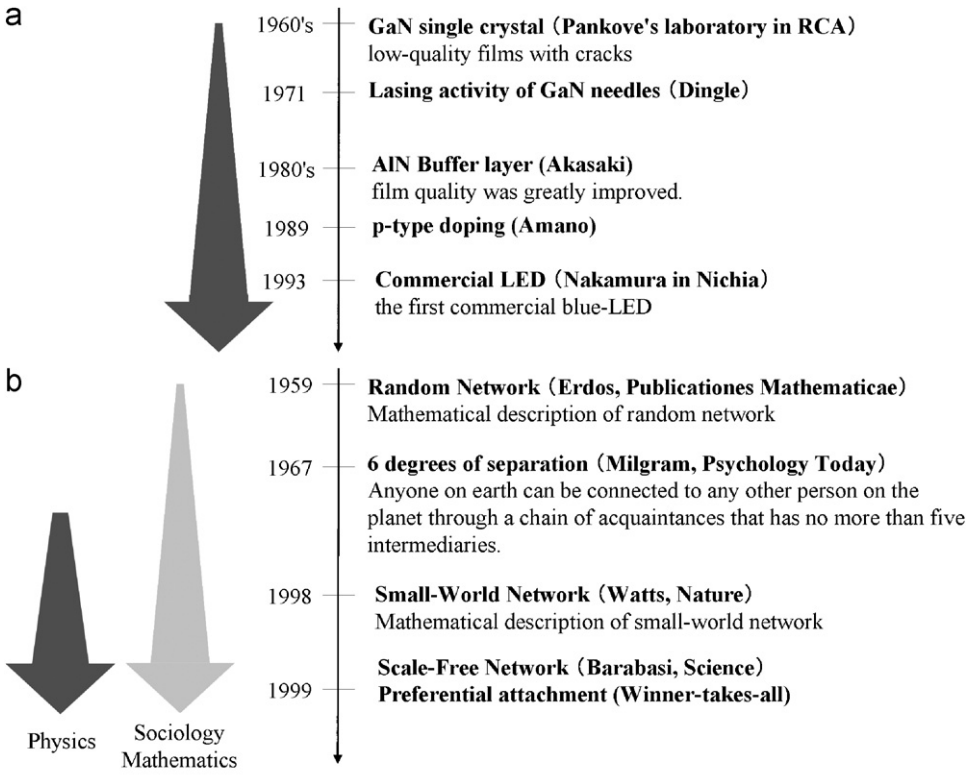

Fig. 1. Brief description of each domain: (a) GaN and (b) complex networks.

Japan (Nakamura et al., 1994). They also succeeded in realizing LD (Nakamura et al., 1996). The rapid development of GaN research is attracting both researchers and funding. The number of academic papers starts to dramatically increase after 1995. In 2004, we annually have about 2000 papers reporting on the material (Fig. 2(a)).

2.2. Complex networks

The second example is the CN, which is also recently recognized as pioneering a new research field. Complex networks have been researched by several types of researchers. Traditionally, the study of CN has been the territory of graph theory in mathematics and also social network analysis in sociology. In the 1950s, two mathematicians, Paul Erdös and Alfréd Rényi, proposed a random graph, which is the simplest and most straightforward example of CN (Erdös and Rényi, 1959, 1960, 1961). The formal elegance of the random model has arrested many mathematicians. On the other hand, in sociology, applied research has studied a variety of networks from individuals to families and nations. One well-known instance is the "six degrees of separation" theory by the social psychologist, Milgram (1967). His famous theory claims that there is a path of acquaintance with a typical length of about six between most pairs of people in the United States. Milgram's accomplishment was to show empirical evidence of the daily wonder of our "small world."

Recently, Watts and Barabási, whose backgrounds were theoretical and applied mechanics and applied physics revealed the common characteristics of small-world networks (Watts and Strogatz, 1998) and scale-free networks (Barabási and Albert, 1999). After the leading works by these physicists, studies in this domain have received attention, and a number of papers in this domain have been published. This is probably because their superiority in expressing small-world phenomena by concise quantitative measures assures researchers of future possibility in this research domain. Although even in 1977 a sociologist, Freeman (1977), quantitatively treated networks and proposed the concept of centrality, most sociologists' work still comprises qualitative and descriptive reports. Currently, we have more than 1000 papers annually (Fig. 2(b)).

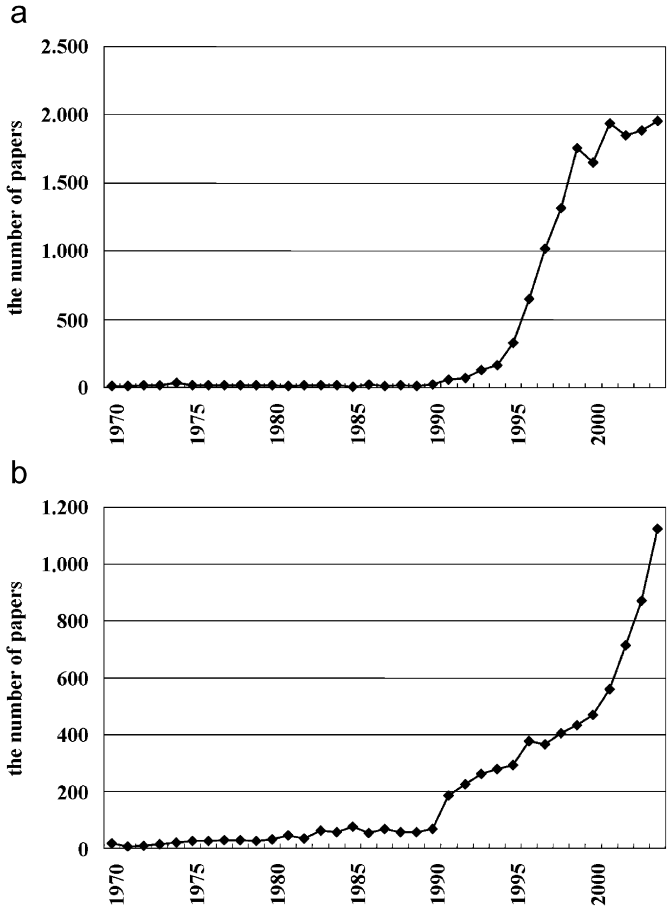

Fig. 2. Changes in the number of papers: (a) GaN and (b) complex networks.

3. Research methodology

In this section, the methodology of this research is shown. Analyzing schema is depicted in Fig. 3. The first step is to collect the data of each knowledge domain and to make citation networks for each year. The problem, how we should define a research domain, is difficult to solve. One solution is to use a keyword that seems to represent the research domain. When we collect papers retrieved by the keyword, we can make the corpus for the research domain. However, it causes two problems. One is the deficiency of relevant papers. It is not always true that a research domain can be represented by a single keyword. Another is the surplus of papers. In some cases, the same keyword is used in different research domains, which includes the noisy papers to the corpus. To overcome the first problem, we use broad queries to retain wide coverage of citation data. For the second problem, we analyze only the maximum component of the citation networks. (A component is an "isolated" part of a citation network, which does not have citations to and from another part, and the maximum component is the part that includes most papers in it.) By doing this step, non-relevant papers that do not cite papers in the corresponding research domain are removed. With this process, proper papers can be necessarily in the maximum component, because proper papers cite or are cited by certain important papers, which gain many citations and are central in maximum component.

After extracting the maximum component, we perform the topological clustering, in order to discover tightly knit clusters with a high density of within-cluster edges with Newman's algorithm (Newman, 2004). With this process, citation networks are divided into clusters, within which papers cite densely each other. In the last step, two topological measures, within-cluster degree, z_i, and

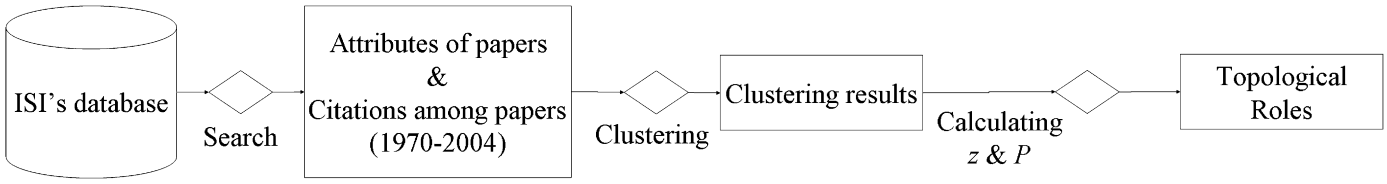

Fig. 3. Method proposed in this paper.

participation coefficient, P_i, proposed by Guimera and Amaral (2005) are calculated in order to track the position of each paper in the clustered citation network.

3.1. Data collection

We collect citation data from the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI) compiled by the Institute for Scientific Information (ISI), which maintains citation databases covering thousands of academic journals and offers bibliographic database services, because SCI and SSCI are two of the best sources for citation data. We use Web of Science, which is a web-based user interface of ISI's citation databases. Papers published after 1970 are contained in ISI's citation databases. We search the papers using the following terms as queries: "GaN OR Gallium Nitride" for the first domain, and "social networks OR social network OR random networks OR random network OR small-world OR scale-free OR complex networks" for the second domain. In our method, the queries are selected as the following two steps: (1) the representative keyword, such as GaN and social network, was selected; (2) if the definition of its domain is unclear, more keywords, such as random network, small-world, scale-free, and CN, should be added. The second step is known as query expansion (Kostoff et al., 1997). Our intention to use such many terms is to retain wide coverage of citation data in order to avoid the omission of significant papers. As a result, we obtained the data of 15,134 papers on GaN and 7370 papers on CN that had been published from 1970 to 2004. Fig. 2(a) shows the changes over time in the number of papers in GaN, and Fig. 2(b) shows the changes over time in CN. In the area of GaN, the number of papers started increasing dramatically in 1995 and reached a peak in 2000, whereas in CN, the number of papers started growing in 2000 and is currently increasing.

The ISI's citation databases enable us to obtain both the attribute data of each paper such as the year published, title, author(s), abstract, and so on, and relational data, i.e., citation data. We create citation networks by regarding papers as nodes and intercitations as links. The network created in each year enables a time-series analysis of citation networks. In traditional citation analysis, the citation threshold is usually set to reduce the calculation load, and the citation patterns derived therefore reflect the mainstream domain knowledge. The results of the pruning papers by citation threshold often consist of chains among highly cited classic works. It excludes relatively infrequently cited works such as recent contributions from the analysis (Chen et al., 2001). The use of co-citation in traditional citation analysis is also problematic (Hopcroft et al., 2004). In co-citation, two papers are linked if they are both cited by another paper published later. This is a useful similarity measure. However, in order to make this measure work properly, a certain time lag is inescapably needed in order that papers build up a citation record when we use co-citation. Suppose a following case, shown in Fig. 4.

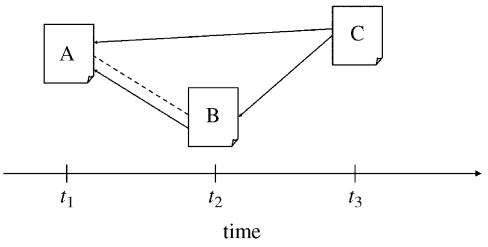

Fig. 4. Difference between intercitation and co-citation. Solid line is an intercitation and dotted line is a co-citation.

(1) At t_1, node A is born.
(2) At t_2 ($>t_1$), node B is born and B cites A.
(3) At t_3 ($>t_2$), node C is born and C cites both A and B.

When we use intercitation analysis, edge from B to A is made at t_2. But in co-citation analysis, such a link is not formed immediately, because A and B are not cited by the same paper when B is published at t_2. If we use co-citation analysis, the link between A and B is formed only after C citing both A and B is published at t_3. In sum, although with co-citation analysis edge between B and C is not made until t_3, with intercitation analysis edge from B to A is made at t_2. Intercitation analysis is more sensitive to recent citations than co-citation analysis. Because our objective is to detect changes as early as possible, we opted for intercitation that uses the common reference set of each paper. Additionally, the analysis of intercitation is more straightforward than co-citation. Klavans and Boyack (2006) compared the similarity of the clustering results by intercitation to that by co-citation. They concluded that intercitation is more appropriate for the clustering of the similar documents. Intercitation also allows us to group papers that are only rarely cited, which is a significant

portion of all papers (Hopcroft et al., 2004). In network analysis, only the data of largest-graph component is used, because this paper focuses on the relationship among papers, and we should therefore eliminate the papers that have no link with any other papers. The number of papers contained in the largest component in 2004 is 14,240 (94% of collected papers) in GaN and 3524 (48%) in CN. The reason why the size of the largest component in CN is relatively small was because we used more queries. We used many queries in order to avoid the omission of significant papers.

3.2. Statistical method

In this work, we analyze intercitation networks of scientific publications in the above two domains with topological clustering and extracting topological measures. The topological clustering divides papers into some clusters and the topological measures are determined after the clustering. We focus on the topological clustering method in order to discover tightly knit clusters with a high density of within-cluster edges, which enables the creation of a non-weighted graph consisting of a large number of nodes. Citation networks where each paper is connected by intercitation are divided into clusters. Clustering of a citation network divides collected papers having similar citations into the same cluster in order to specify the research domain. Although these clustering methods have been difficult to achieve due to the difficulty in cluster analysis of non-weighted graphs consisting of a large number of nodes, there has in recent years been a lot of progress in methods for topological clustering with great progress in CN research (Newman, 2004). After clustering, the position of each paper in the clustered citation network is tracked using topological measures, i.e., within-cluster degree, z_i, and participation coefficient, P_i, proposed by Guimera and Amaral (2005).

3.2.1. Clustering

Amongst many clustering methods and algorithms, in this paper we apply a method proposed by Newman which is able to deal with large networks with relatively small calculation time in the order of $O((m+n)n)$, or $O(n^2)$ on a sparse network, with m edges and n nodes; therefore, this could be applied to large-scale networks (Newman, 2004). The algorithm proposed is based on the idea of modularity. Modularity Q was defined as follows:

$$Q = \sum_s (e_{ss} - a_s^2) = Tr(e) - ||e||^2 \quad (1)$$

where e_{st} is the fraction of the edges in the network that connect nodes in cluster s to those in cluster t, and $a_s = \sum_t e_{st}$. The first part of the equation, $Tr(e)$, represents the sum of density of edges within each cluster. A high value of this parameter means that nodes are densely connected within each cluster. However, the maximum value of this $(Tr(e) = 1)$ is given if whole nodes are regarded as one cluster. The second part of the equation, $||e||^2$, represents the sum of density of edges within each cluster when all edges are placed randomly. For instance, suppose that all nodes are regarded as one cluster. In this case, matrix e has only one row, $e = (1)$. Therefore, $Tr(e) = 1$, $||e||^2 = 0$, and $Q = 1$, which is the theoretical maximum value of Q. That is, Q is the fraction of edges that fall within communities, minus the expected value of the same quantity if the edges fall at random without regard for the community structure. Newman's method cuts off edges that connect clusters sparsely and extract clusters within which nodes are connected densely. A high value of Q represents a good community division where only dense-edged remain within clusters and sparse edges between clusters are cut off, and $Q = 0$ means that a particular division gives no more within-community edges than would be expected by random chance. Then, the algorithm to optimize Q over all possible divisions to find the best structure of clusters is as follows. Starting with a state in which each node is the only member of one of n clusters, we repeatedly join clusters together in pairs, choosing at each step the join that results in the greatest increase in Q. The change in Q upon joining the two clusters is given by

$$\Delta Q = e_{st} + e_{ts} - 2a_s a_t = 2(e_{st} - a_s a_t) \quad (2)$$

In this paper, we stop joining when ΔQ became minus, because the purpose here is not to gain a whole dendrogram but extract more relevant structures with regard to the citation networks.

3.2.2. Extracting the role of each paper

After dividing the papers into optimized clusters using Newman's method, the role of each paper is determined by its within-cluster degree and its participation coefficient, which define how the node is positioned in its own cluster and between clusters (Guimera and Amaral, 2005). This method is based on the idea that nodes with the same role should be at similar topological positions. These two properties can be easily calculated after clustering the network. Within-cluster degree z_i measures how "well connected" node i is to other nodes in the cluster, and is defined as

$$z_i = \frac{\kappa_i - \overline{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}} \quad (3)$$

where κ_i is the number of edges of node i to other nodes in its cluster s_i, $\overline{\kappa}_{s_i}$ is the average of κ over all nodes in s_i, and $\sigma_{\kappa_{s_i}}$ is the standard deviation of κ in s_i. z_i is high if the within-cluster degree is high and vice versa.

Participation coefficient P_i measures how "well distributed" the edges of node i are among different clusters and is defined as

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{k_i}\right)^2 \quad (4)$$

where κ_{is} is the number of edges of node i to nodes in cluster s, and k_i is the total degree of node i (the number of edges of node i). Participation coefficient P_i is close to 1 if its edges are uniformly distributed among all the clusters and 0 if all its edges are within its own cluster.

Guimerà and Amaral applied this analysis to biological networks and heuristically defined seven different universal roles, by a different region in the $z-P$ parameter space as shown in Fig. 5. According to the within-cluster degree, they classified nodes with $z \geqslant 2.5$ as hub nodes and nodes with $z < 2.5$ as non-hub nodes. In addition, non-hub nodes can be naturally divided into four different roles: (R1) ultra-peripheral nodes; that is, nodes with most of their edges within their cluster ($P < 0.05$), (R2) peripheral nodes; that is, nodes with many edges within their cluster ($0.05 < P \leqslant 0.62$); (R3) non-hub connector nodes, that is, nodes with a high proportion of edges to other clusters ($0.62 < P \leqslant 0.80$); and (R4) non-hub kinless nodes, that is, nodes with edges homogeneously distributed among all clusters ($P > 0.80$). Similarly, hub nodes can be classified into three different roles: (R5) provincial hubs, that is, hub nodes with the vast majority of edges within their cluster ($P < 0.30$); (R6) connector hubs, that is, hubs with many edges to the other clusters ($0.30 < P \leqslant 0.75$); and (R7) kinless hubs, that is, hubs with edges homogeneously distributed among other clusters ($P > 0.75$).

3.2.3. Topic detection by Natural Language Processing

In this section, the method of extracting the characteristic terms for each cluster by Natural Language Processing (NLP), which enable research topic detection, is described. First of all, candidate terms are extracted by linguistic filtering, using all abstracts of papers (Mima et al., 1998; Frantzi et al., 2000). Linguistic filtering extracts candidate noun phrases, such as

(1) *Noun*+*Noun*,
(2) (*Adj*|*Noun*)+*Noun*,
(3) ((*Adj*|*Noun*)+|((*Adj*|*Noun*) (*NounPrep*)?)(*Adj*|*Noun*))*Noun*.

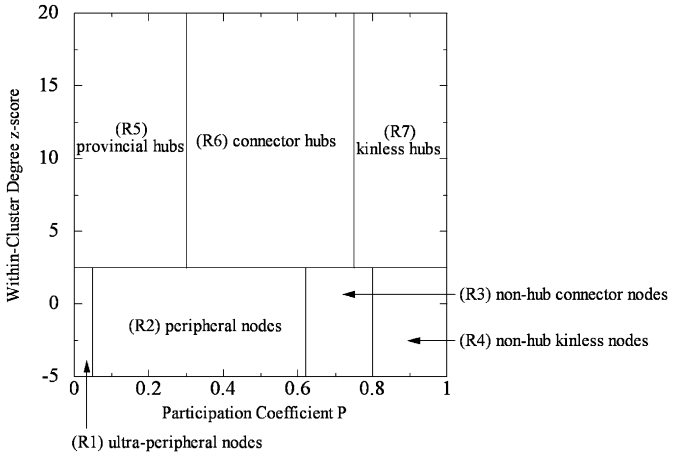

Fig. 5. Role of each node in the topology.

Then, these noun phrases are weighted by *tf–idf* weight, which is a weight often used in information retrieval. The term frequency, *tf*, in the given documents gives a measure of the importance of the term within the particular document. The inverse document frequency, *idf*, is a measure of the general importance of the term, which is the log of the number of all documents divided by the number of documents containing the term, enabling common terms to be filtered out. Therefore, a term with high *tf–idf* means that the term has a high term frequency in the given document and a low document frequency of the term in the entire documents. The *tf–idf* weight of term i in document j is given by

$$w_{i,j} = tf_{i,j} \times idf_i = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{5}$$

where $tf_{i,j}$ is the number of occurrences of term i in document j, $idf_i = \log(N/df)$ is the inverse document frequency, a measure of the general importance of the term, df_i is the number of documents containing term i, and N is the total number of documents. In this paper, in order to extract the important terms not in a certain document but in a certain cluster, we extend the *tf–idf* weight to clusters, and the *tf–idf* weight of term i in cluster s is given by

$$w_{i,s} = tf_{i,s} \times idf_i = tf_{i,s} \times \log\left(\frac{N}{df_s}\right) \tag{6}$$

where $tf_{i,s}$ is the number of occurrences of term i in cluster s. In this paper, the top ten terms of the *tf–idf* (term frequency–inverse document frequency) value in each cluster are regarded as terms characteristic of that cluster.

4. Results

Fig. 6(a) shows the number of papers in each cluster and the average age of the papers included in each cluster from 1992 to 2004 in GaN. As shown in Fig. 6(a), the age of the clusters, which is the average of the ages of the papers in the cluster, seems to decrease up to 1998. This is because many fresh papers join these clusters, therefore reducing the ages of these clusters. However, after 1998, the ages of large clusters increase, which shows that these research domains have become mature. In 2004, there are three big clusters that we name G_1, G_2, and G_3. In 2004, G_1 includes 2509 papers and is 4.4-years old, which is older than G_2 (2267 papers and 1.8-years old), and G_3 (1525 papers and 1.4-years old).

Fig. 6(a) shows the historical succession of the papers included in these clusters. The similarity of papers included in the cluster is the highest in the G_1 cluster. In other words, most of the papers in G_1 in a certain year will also appear in G_1 in subsequent years. On the other hand, G_2 and G_3 are strongly mixed and interrelated. Table 1 shows a list of the most cited papers in these clusters. G_1 includes breakthrough papers reporting new processes such as the introduction of a buffer layer and a successful p-type

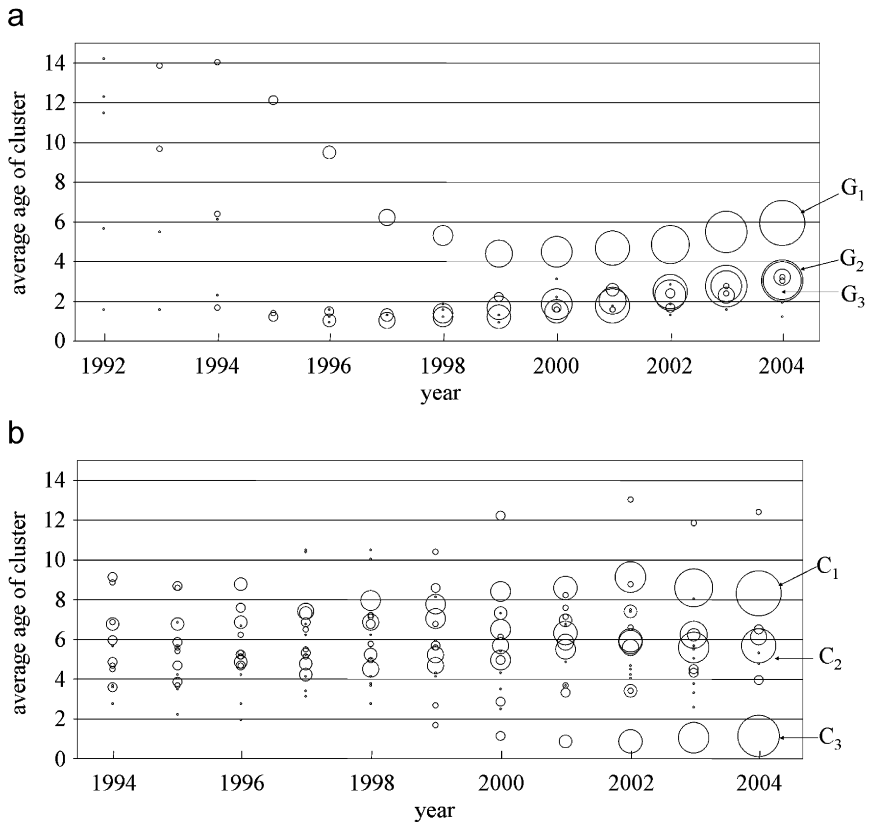

Fig. 6. Cluster size and average age. The circles are clusters, and the size of each circle means the relative value of the number of papers in each cluster in (a) GaN and (b) complex networks.

doping technique, which are the key to realizing blue LEDs, while research in G_2 and G_3 focus on the study of the physical properties of synthesized GaN films and the development of light-emitting devices. Moreover, the average age of papers in G_1 is still higher than the other clusters. These results suggest that innovation in GaN mainly occurs in the traditional research domain of process development, G_1, rather than in newly developing research on physical properties and device fabrications. This result accords with the fact that the breakthrough in GaN domains is recognized to occur in new process development such as the buffer layer and *p*-type doping. In a certain research field, tracking the average age of each cluster in the domain is effective in detecting emerging research fronts where innovative breakthrough occurs. However, we must note that such breakthrough appears in the traditional cluster; therefore, we call this type of innovation incremental innovation. As shown in Fig. 7, in GaN each cluster was strongly connected and these connections contributed the growth of existing clusters, rather than the creation of new clusters.

Fig. 6(b) shows the number of papers in each cluster and average age of the papers included in each cluster from 1994 to 2004 in CN. There is a different tendency between GaN and CN. In CN, the age of most of the clusters seems to be unchanged, which indicates the gradual development of research in this domain. However, in CN, there seems to be a new emerging cluster in 2000 of the age of about one as seen in the bottom of Fig. 6(b). This is also confirmed by the historical succession of the papers included in these clusters (Fig. 7(b)). In 2004, there are three main clusters, which we name C_1, C_2, and C_3 clusters. C_1 has 1256 papers and is 8.3-years old, C_2 has 785 papers and is 5.7-years old, and C_3 has 1099 papers and is 1.1-years old. Table 2 shows a list of the most cited papers in these clusters in 2004. C_1 and C_2 typically have sociological origins, while papers in C_3 are published in physics journals. This shows that in CN, physical clusters appear from traditional sociological research domains, but not in those domains, and form a new research domain. We call this type of innovation branching innovation.

Fig. 8 shows plots of z and P for the top 10 papers in the number of citations of 2000 in GaN and of 2004 in CN. These scores have year-to-year variation. There is a remarkable difference between these two types of innovation, incremental, and branching innovations. In GaN, where incremental innovation occurred, the top 10 papers changed position from (R2) peripheral nodes to (R6) connector hubs as the domain developed. However, in CN, where branching innovation occurred, the top 10 papers moved from (R1) ultra-peripheral nodes to (R5) provincial hubs, and became provincial hubs. This means that there

Table 1
Clustering result of GaN at 2000. TC(t) means the number of citations and in year t

Cluster id	# Paper	Average age	Top 10 *tf–idf* terms	Papers (TC (2000)⩾150	Year	TC (2000)
G_1	2509	4.4	Degree, growth, substrate, films, ga, gaas, gan, nh, si, surface	Strife, S., 1992, Journal of Vacuum Science and Technology B 10, 1237	1992	837
				Nakamura, S., 1994, Applied Physics Letters, 64, 1687	1994	659
				Amano, H., 1989, Japanese Journal of Applied Physics Part 2, 28, L2112	1989	519
				Amano, H., 1986, Applied Physics Letters, 48, 353	1986	488
				Nakamura, S., 1991, Japanese Journal of Applied Physics Part 2, 30, L1705	1991	395
				Akassaki, I., 1989, Journal of Crystal Growth, 98, 209	1989	306
				Nakamura, S., 1992, Japanese Journal of Applied Physics Part 1, 31, 1258	1992	295
				Dingle, R., 1971, Physical Review B, 4, 1211	1971	293
				Strite, S., 1991, Journal of Vacuum Science and Technology B, 9, 1924	1991	217
				Paisley, M.J., 1989, Journal of Vacuum Science and Technology A, 7, 701	1989	213
				Nakamura, S., 1995, Japanese Journal of Applied Physics Part 2, 34, L 797	1995	198
				Monemar, B., 1974, Physical Review B, 10, 676	1974	197
				Powell, R.C., 1993, Journal of Applied Physics, 73, 189	1993	181
				LEI T, 1991, Applied Physics Letters, V59, P944	1991	172
				Nakamura, S., 1992, Japanese Journal of Applied Physics Part 2, 31, L 139	1992	163
				Davis, R.F., 1991, Proceedings of the IEEE, 79, 702	1991	163
				Dingle, R., 1971, Solid State Communications, 9, 175	1971	160
				Lei, T., 1992, Journal of Applied Physics, 71, 4933	1992	156
G_2	2267	1.8	Degree, contact, gan, al, ni, ga, ti, au, physics, american institute	Morkoc, H., 1994, Journal of Applied Physics, 76, 1363	1994	518
				Mohammad, S.N., 1995, Proceedings of the IEEE, 83, 1306	1995	206
				Neugebauer, J., 1994, Physical Review B, 50, 8067	1994	205
				Ogino, T., 1980, Japanese Journal of Applied Physics, 19, 2395	1980	181
				Barker, A.S., 1973, Physical Review B, 7, 743	1973	176
				Chichibu, S., 1996, Applied Physics Letters, 69, 4188	1996	172
				Bernardini, F., 1997, Physical Review B, 56, 10024	1997	166
G_3	1525	1.4	gan, mg, layers, ga, physics, american institute, structures, defects, photoluminescence, strain	Nakamura, S., 1996, Japanese Journal of Applied Physics Part 2, 35, L74	1996	500
				Lester, S.D., 1995, Applied Physics Letters, 66, 1249	1995	285
				Nakamura, S., 1995, Japanese Journal of Applied Physics Part 2, 34, L1332	1995	236
				Nakamura S, 1998, Applied Physics Letters, 72, 211	1998	180
				Akasaki, I., 1996, Electronics Letters, 32, 1105	1996	167
				Usui, A., 1997, Japanese Journal of Applied Physics Part 2, 36, L899	1997	160

was only a narrow bridge among clusters such as the ''Physics'' cluster and the ''Social'' cluster. In other words, they became independent clusters as the research developed. Furthermore, in CN, existing papers, such as BERKMAN LF, 1979, and FREEMAN LC, 1979, were at (R6) connector hubs before C_3 appeared as an independent cluster. This result revealed that although sociological papers were globally central before physics papers started to be published, these became not connector hubs but provincial hubs after the activated center of research shifted from sociology to physics.

The results also supported the visualization of citation networks and topic detection by NLP. In order to visualize citation maps, we used large graph layout (LGL), an

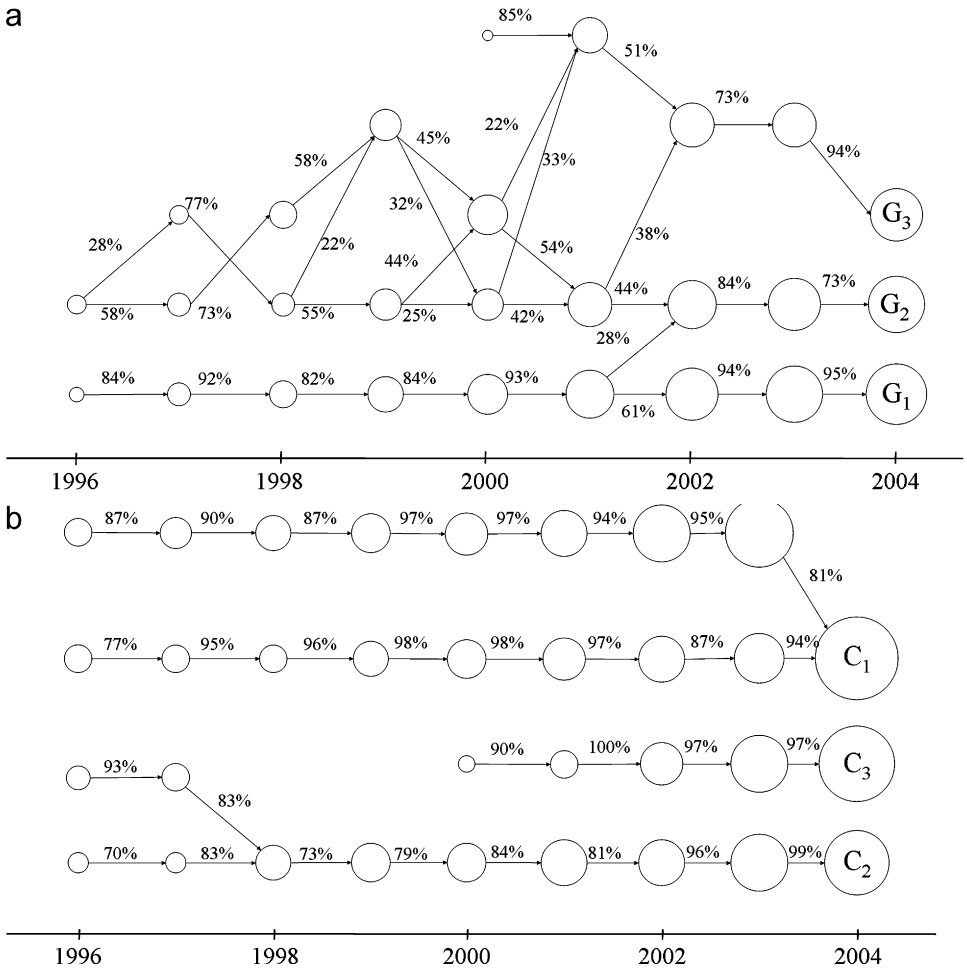

Fig. 7. Time line visualization of the development of each cluster in (a) GaN and (b) complex networks. The circles are clusters, and the size of each circle means the relative value of the number of papers in each cluster. The percentages from cluster i in year t to j at $t+1$ in this figure mean (the number of papers from cluster i at t to j at $(t+1)$)/(the number of papers in cluster i at t).

algorithm developed by Adai et al. (2004), which can be used to dynamically visualize large networks in the order of hundreds of thousands of nodes and millions of edges, and applies a force-directed iterative layout guided by a minimal spanning tree of the network in order to generate coordinates for the nodes in two or three dimensions. We visualize the citation network by expressing intercluster links as the same color. Through this visualization, clusters are intuitively understood. Characteristic terms for each cluster extracted by NLP are shown in Table 3, and the visualization of the clustering result of CN in 2000 is shown in Fig. 9. Characteristic terms for each cluster were automatically extracted by NLP from abstracts of papers, and only the cluster names in Fig. 9 were manually assigned. For example, in the "Social, Support or Disease" cluster, patients, support, depression, schizophrenia, clients, mental illness, social support, and so on were discussed. The average age of papers in the "Social, Support or Disease" cluster was 8.4-years old. There were also some other large clusters such as "Social, Network Analysis" cluster which was 6.5-years old and in which social structure were discussed, "Social, Support" cluster which was 4.9-years old and about supports, health, association, smoking and survival, and "Social, HIV" cluster which was a rather young cluster and about infections through social networks. Other than that, the papers in the "Physics, Small-World "cluster in 2000, whose average age was a mere 1.1, apparently had different research topics from other domains because these papers were published in physical journals such as *Science*, *Nature*, and *Physical Review*, whereas papers of other clusters were mainly published in sociological journals.

Fig. 10 is a time series of a citation network in GaN for Fig. 10(a) and CN for Fig. 10(b-1) and (b-2) visualized by LGL. In the visualization, the positions of all the papers are fixed at the position calculated by using the data from 1970 to 2004 in Fig. 10(a) and Fig. 10(b-1), whereas in Fig. 10 (b-2) positions are with the data from 1970 to each year. Each link is shown when each paper including the link is published. The links with the same color belong to the same cluster. Looking at Fig. 10(b-1), we can see three independent clusters; sociological clusters, a "Physics,

Table 2
Clustering result of complex networks at 2004

Cluster id	# Papers	Average age	Top 10 *tf–idf* terms	Papers (TC (2004)⩾50)	Year	TC (2004)
C_1	1256	8.3	Support, women, patients, men, health, age, social support, friends, studies, loneliness, mortality	Berkman, L.F., 1979, American Journal of Epidemiology, 109, 186	179	252
				Tolsdorf, C.C., 1976, Family Process, 15, 407	1976	110
				Orthgomer, K., 1987, Journal of Chronic Diseases, 40, 949	1987	55
				Mckinlay, J.B., 1973, Social Forces, 51, 275	1973	54
				Seeman, T.E., 1998, Social Science and Medicine, 26, 737	1988	51
C_2	785	5.7	Model, women, children, groups, patients, paper, studies, structure, families, developments, article	Freeman, L.C., 1979, Social Networks, 1, 215	1979	162
				Klovdahl, A.S., 1985, Social Science and Medicine, 21, 1203	1985	67
C_3	1099	1.1	Nodes, scale, graphs, model, vertices, proteins, links, distribution, topologies, degree distribution, connectivity	Watts, D.J., 1998, Nature, 393, 440	1998	722
				Barabasi, A.L., 1999, Science, 286, 509	1999	558
				Albert, R., 2002, Review of Modern Physics, 74, 47	2002	499
				Strongatz, S.H., 2001, Nature, 410, 268	2001	299
				Albert, R., 2000, Nature, 406, 268	2000	248
				Jeong, H., 2000, Nature, 407, 651	2000	243
				Dorogovtsev, S.N., 2002, Advanced Physics, 51, 1079	2002	210
				Barabasi, A.L., 1999, Physica A, 272, 173	1999	148
				Newman, M.E.J., 2003, SIAM Review, 45, 167	2003	133
				Cohen, R., 2000, Physics Review Letters, 85V, 4626	2000	119
				Newman, M.E.J., 1999, Physics Review E, 60, 7332	1999	112
				Barrat, A., 2000, European Physics Journal B, 13, 547	2000	106
				Krapivsky, P.L., 2000, Physics Review Letters, 85, 4629	2000	104
				Liljeros, F., 2001, Nature, 411, 907	2001	103
				Callaway, D.S., 2000, Physics Review Letters, 85, 5468	2000	99
				Barthelemy, M., 1999, Physics Review Letters, 82, 3180	1999	90
				Newman, M.E.J., 2002, Physics Review Letters, 89	2002	84
				Moore, C., 2000, Physics Review E, 61, 5678	2000	83
				Ravasz, E., Science, 297, 1551	2002	78
				Cohen, R., 2001, Physics Review Letters, 86, 3682	2001	71
				Kuperman, M., 2001, Physics Review Letters, 86, 2909	2001	62
				Newman, M.E.J., 1999, Physics Letters A, 263, 341	1999	61
				Milo, R., 2002, Science, 298, 824	2002	60
				Newman, M.E.J., 2000, Physics Review Letters, 84, 3201	2000	59
				Kleinberg, J.M., 2000, Nature, 406, 845	2000	51
				Barabasi, A.L., 2000, Physica A, 281, 69	2000	51
				Monasson, R., 1999, European Physics Journal B, 12, 555	1999	50

Water'' cluster, and a ''Physics, Small-World'' cluster. Comparing 1998 with 2004, we can see that in 1998, only the sociological cluster was visible on the upper left; however, in 2004, papers by physicists appeared on the lower right, while sociological papers were continuously published. In fact, after about 2001, papers by physicists became detectable in Fig. 10(b-1). Fig. 9 is an enlargement of a visualized citation map in 2000 in Fig. 10(b-2). We can also detect a rather small cluster as a branch of ''Social, Network Analysis'' cluster. The interdependence of these clusters is clearly seen in Fig. 10(b-2) where we use the citation data at each year and successively visualize these data. When we visualize citation network at each year, we can see that ''Physics, Small-World'' cluster was born as a branch of a traditional cluster and then gradually became an independent cluster. Therefore, we can interpret and conclude it as physical clusters emerge from traditional sociological research domains but not in those domains

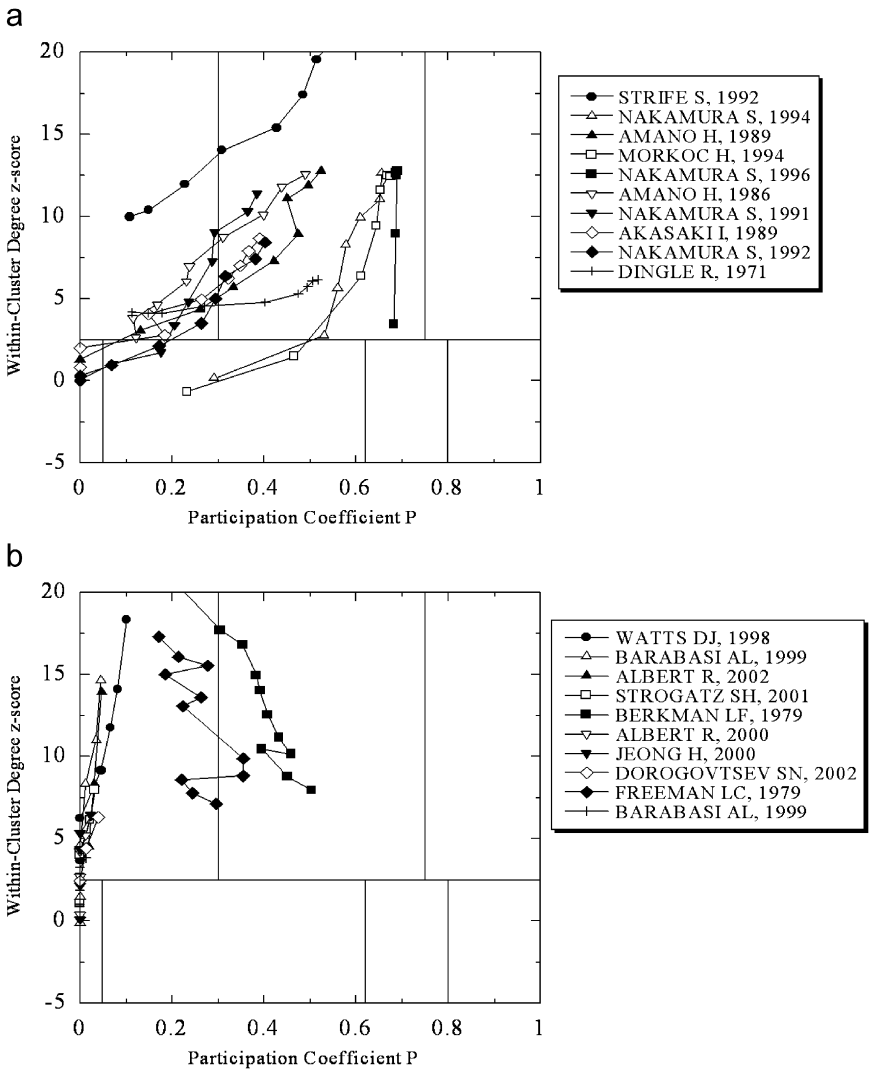

Fig. 8. Changes in the roles of top 10 papers of times cited in (a) GaN and (b) complex networks.

and form a new research domain. These visualizations enable us to understand when these clusters emerged visibly. Fig. 10(a) is a time series of a citation network in GaN. In contrast to CN, in GaN, each cluster is difficult to distinguish, and the entire papers incrementally increased as one strongly connected body. As we showed by topological measures, differences between incremental and branching innovations are also visible with visualizations.

5. Discussions

As described above, we performed a comparative study in two research domains to develop a method of detecting emerging research domains. We performed citation network analysis on GaN and CN research. Papers dealing with a similar topic cite each other and are strongly connected, and papers dealing with different topics are weakly connected. Therefore, the division of a knowledge domain into strongly connected clusters is necessary in order to detect emergence. We performed topological clustering method to detect such emerging research clusters, and then track the topological positions of each paper by z and P.

By performing a comparative study, we found that two types of innovation, incremental innovation and branching innovation, can be distinguished by our method. In incremental innovation, breakthrough occurs and develops within traditional research clusters, and reflecting it hub papers are connector hubs with large z and large P. On the other hand, in branching innovation, breakthrough occurs from traditional research clusters but develops as an independent cluster. In this case, active research centers shift rapidly, and hub papers become provincial hubs with large z and small P. Two types of innovation can be distinguished by using topological measures, i.e., z and P. Therefore, monitoring z and P enable us to judge how each domain developed.

Table 3
Clustering and NLP result of complex networks in 2000

Cluster id	# Papers	Average age	Top 10 *tf–idf* terms	Papers (TC (2000)⩾30)	Year	TC (2000)
C'_1	331	8.4	Patients, supports, depression, schizophrenia, clients, mental illness, social support, women, child, families, treatment	Tolsdorf, C.C., 1976, Family Process, 15, 407	1976	102
				Mckinlay, J.B., 1973, Social Forces, 51, 275	1973	47
				Hirsch, B.J., 1979, American Journal of Community Psychology, 7, 263	1979	39
C'_2	322	6.5	Model, scale, child, patients, women, structure, families, paper, group, development, relationship	Freeman, L.C., 1979, Social Networks, 1, 215	1979	95
				Breiger, R.L., 1975, Journal of Mathematical Psychology, 12, 328	1975	39
C'_3	281	4.9	Women, men, mortality, ci, supports, health, association, age, smoking, year survival	Berkman, L.F., 1979, American Journal of Epidemiology, 109, 186	1979	189
				Orthgomer, K., 1987, Journal of Chronic Diseases, 40, 949	1987	43
				Seeman, T.E., 1987, American Journal of Epidemiology, 126, 714	1987	36
				Hanson, B.S., 1989, American Journal of Epidemiology, 130, 100	1989	30
				Seeman, T.E., 1988, Social Science and Medicine, 26, 737	1988	30
C'_4	71	1.1	World, dynamics, site, connectivity, model, graphs, transition, phenomena, vertices, probability, games	Watts, D.J., 1998, Nature, 393, 440	1998	46
C'_5	71	2.9	HIV, infection, syphilis, risk, HIV infection, drug injectors, transmission, persons, epidemic, HIV transmission, AIDS	Klovdahl, A.S., 1985, Social Science and Medicine, 21, 1203	1985	43
				Klovdahl, A.S., 1994, Social Science and Medicine, 38, 79	1994	32

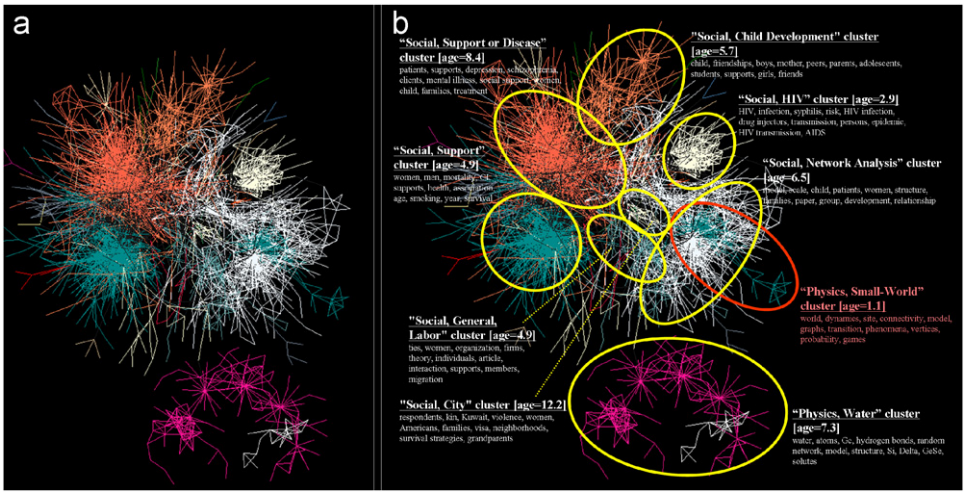

Fig. 9. Citation networks with characteristic terms of complex networks at 2000. The same visualizations are drawn in parts (a) and (b). The position of each citation is calculated by LGL algorithm using the citation data from 1970 to 2000. The same edge colors means the same cluster that was calculated by Newman's modularity Q. The thick yellow line in part (b) is a guide for the eye. In part (b), cluster names and characteristic terms were also shown.

Currently, the management of R&D activity faces increasing difficulty in overviewing diverse research domains and detecting emerging research fronts due to the specialization and segmentation of research domain as well as the flood of information. However, there is still a lack of researchers, R&D managers, and policy makers

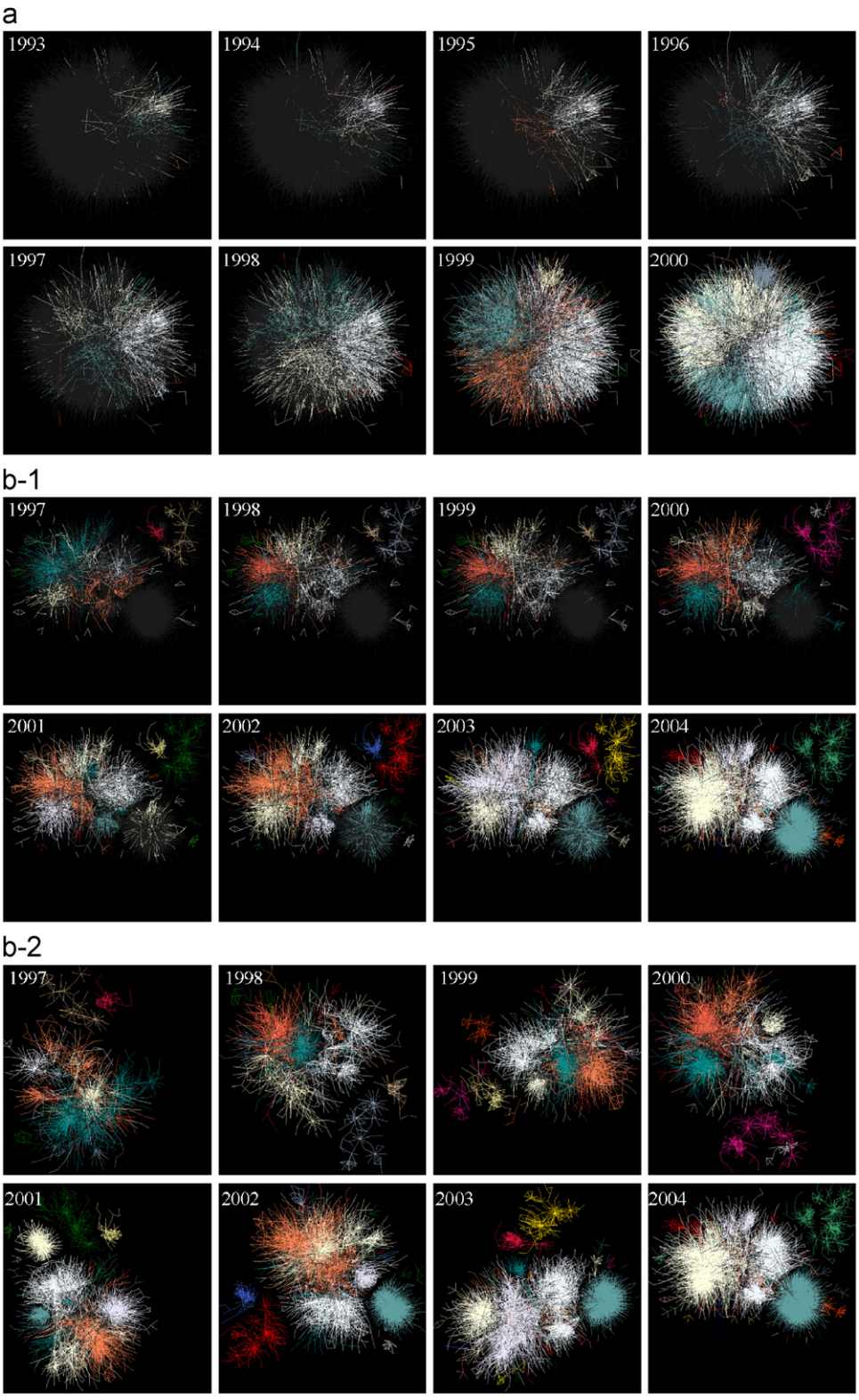

Fig. 10. Visualization of the evolution of citation network in (a) GaN and (b-1, b-2) complex networks. In Fig. 9(a) and (b-1), the position of each citation is calculated by using the citation data from 1970 to 2004 and then fixed in other figures, whereas in Fig. 9 (b-2) the position us calculated by using citation data from 1970 to each year.

overviewing scientific activities and detecting emerging research domains. Our topological approach can become a tool to assist them to detect emerging research domains among a pile of publication in a manner that meet a commensurate increasing need as scientific and technical intelligence to discover emerging research domains in an

era of information flooding. Our approach can also be utilized to describe how innovation occurs in traditional discipline, coevolves within it, and/or become an independent discipline.

Then, let us discuss the limitation of our research concerning with the broadness of our queries we used and their influences. Our intention to use such many terms for CN is to retain wide coverage of citation data in order to avoid the omission of significant papers. These operations did not distort our conclusion. Suppose that we have citation data of all papers. In that case, if we could detect emerging knowledge domains with our method, there is no doubt that "physics cluster" we detected is a newly emerging one. Because collecting all papers in the world is a hard task and not a practical solution, we used broad queries to make the similar corpus to that of whole papers. But then, it is legitimate to have doubts about the influences of query selection on the results. In order to evaluate the influences of query selection, we examined additional calculation using the data collected only by the query "social network*". The results were so similar to our original one that our method could detect the physics papers as a new emerging cluster, independent of broadness of queries. But the year when we can detect a "Physics, Small-World" cluster was 2001, which was 1 year later compared to our original result with the broader queries. It means that we should comprehend a research domain with broader characteristic terms if we want to notice the emerging domain as early as possible.

Finally, let us discuss the significance of our computer-based method, in the viewpoint of contribution to policy making. For both R&D managers in companies or research institutions and policy makers, there are two types of approaches, i.e., expert-based and computer-based approach to notice emerging research domains among numerous academic papers. However, the former approach becomes a highly laborious and difficult task as each research domain becomes specialized and segmented. Our computer-based method, at least, complements this expert-based approach for the following three reasons. First of all, experts' judgment is not always right, especially in the current information-flood era. Sometimes, once-humble researchers accomplish great scientific achievements. Experts may fail to give credit to emerging trends. Second, gathering experts is expensive. Identifying the quality of these papers before they become a new emerging cluster requires numerous experts. Finally, our method is scalable. Even if the publication cycle becomes shorter and the number of publications grows, the computer-based approach could be effective. Moreover, although the previous researches in knowledge mapping emphasized on the method of visualization in order to detect emergence, our method enables us to detect by monitoring variables, such as z and P. When we use visualization, we must judge the emergence of research cluster by the visualized map itself. Utilization of quantitative variables such as z and P open a way to detect it by machine-friendly manner. By tracking the evolution of these variables, we can distinguish the different patterns of innovation as described above.

In this paper, we focused on academic publications and showed that our analyzing schema can be utilized to track the evolution of scientific research, to detect the emerging research domain, and to illustrate the innovation process such as incremental and branching innovation. One direction of future research is to apply it to patent system. The intellectual property rights are becoming significant for the management of firms (Hanel, 2006; Storto, 2006; Bader, 2008). Frietsch and Grupp (2006) illustrated the paradigm shift from bulbs to opto-electronics and photonics by using rather simple science and technology indicators such as the number of publications and patents. By analyzing citation network of academic publications and patents simultaneously, we can more fully understand the process of technical progress.

6. Conclusion

In summary, we performed a comparative study in two research domains to develop a method of detecting emerging research domains. One is a study on GaN, which is widely recognized as a recent prominent innovation in the fields of applied physics and material science. Another is CN analysis, which is also recognized as pioneering a new research field. We divided the papers in each research domain into clusters using the topological clustering method, tracked the evolution of the clusters and the positions of the papers in each cluster, and visualized citation networks with characteristic terms for each cluster.

Papers dealing with a similar topic cite each other and are strongly connected, and papers dealing with different topics are weakly connected. Therefore, the division of a knowledge domain into strongly connected clusters is necessary in order to detect emergence. The method we applied here aims to retain dense connections and remove sparser ones. We analyzed the clustering results using the average age and the historical relation of each cluster.

In the case of GaN, the age of the cluster whose research topic is new process development abruptly decreases up to 1999, which suggests the existence of breakthrough in this cluster. In the case of CN, new cluster by physicists appears from sociological clusters and can be detected in 2000 as a new emerging research front. There are two types of innovation: incremental innovation and branching innovation. In incremental innovation, breakthrough occurs and develops within traditional research clusters. On the other hand, in branching innovation, breakthrough occurs from traditional research clusters but develops as an independent cluster. These two types of innovations can be distinguished by using topological measures, i.e., the within-cluster degree and the participation coefficient. In domains where incremental innovation occurs, hub papers are connector hubs with large z and large P. On the other hand, in the case of branching innovation, there is a new

emerging cluster and active research centers shift rapidly and hub papers become provincial hubs with large z and small P. This means that in the case of GaN, hub papers have intercluster edges, which connect some clusters; however, in the case of CN, hubs connect mainly in their own clusters and have few intercluster edges.

Therefore, monitoring z and P enable us to judge how each domain developed. We also describe the development of each research domain by the visualization of citation networks and topic detection by Natural Language Processing. Our approach, detecting emergence by topological measures, succeeds in distinguishing the type of innovation and noting whether there is an emerging knowledge cluster. Some may complain that monitoring increases in the number of papers in each cluster, without z and P monitoring, is sufficient to detect emergence. However, clustering results change year by year because topology changes, and clusters to which a certain paper belongs differ as time goes by due to the autopoietic nature of citation networks. Autopoietic system is self-creating from their internal interactions, self-organizing, and self-defining of their own boundaries. When we regard academic publications as a system, we can also regard it as an autopoietic one, because it includes internal interactions via mutual citations among papers, and it is self-organizing (past paper determines the future direction of research), and self-defining (scientific domain is determined as it evolves as the publications itself not beforehand). Therefore, monitoring z and P is more beneficial in detecting the emergence than monitoring increases of the number of papers in each cluster, in order to track the development of research domains.

In this paper, we showed that our analyzing schema can be utilized to track the evolution of scientific research, to detect the emerging research domain, and to illustrate the innovation process such as incremental and branching innovation. As shown above, the results supported that the method proposed in this paper could be a computational tool to detect research fronts by using the topological measures of citation networks in addition to visualization. With this method, we could monitor the research fronts and detect the emerging research just by computational calculation. Currently, the management of R&D activity faces increasing difficulty in overviewing diverse research domains and detecting emerging research fronts due to the specialization and segmentation of research domain as well as the flood of information. However, there is still a lack of researchers, R&D managers, and policy makers overviewing scientific activities and detecting emerging research domains. Our topological approach can become a tool for future "Research on Research" (R on R) and can meet a commensurate increasing need as scientific and technical intelligence to discover emerging research domains in an era of information flooding. Our research could promote quantitative method in R on R and technology and innovation management (TIM), and therefore contribute these research domains.

But we must remind the limitation of our approach. In the detection of emerging research domains, the shortcoming of this approach is the existence of time lag. It takes 1 or 2 years until a paper receives citations from other papers. It also takes 1 or 2 years from the completion of research to the publication of the research. Therefore, in the context of TIM and research policy, policy makers should complement this approach with not-published information such as academic conference and expert opinion. We must also notice that we focused only on academic publications in this paper. But it is not sufficient to focus on them when we study innovation. One of the future research directions is to apply our approach to patent system coupled with academic publications. Such an approach will quantitatively describe whether the focal innovation is strongly coupled with traditional research domain or form a new independent one, and how it has been across the academia–industry relations.

Acknowledgments

We thank Hideki Mima, at the Center of Innovation in Engineering Education, School of Engineering, The University of Tokyo, for his help in our research.

References

Adai, A.T., Date, S.V., Wieland, S., Marcotte, E.M., 2004. LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks. Journal of Molecular Biology 340 (1), 179–190.

Adams, J., 1990. Fundamental stocks of knowledge and productivity growth. Journal of Political Economy 98, 673–702.

Akasaki, I., 1998. Evolution of nitride semiconductors. Materials Research Society Symposium Proceedings 482, 3–14.

Akasaki, I., Amano, H., Koide, Y., Hiramatsu, K., Sawaki, N., 1989. Effects of AIN buffer layer on crystallographic structure and on electrical and optical-properties of GaN and Ga1-Xalxn (0-less-than-X-less-than-or-equal-to-0.4) films Grown on sapphire substrate by MOVPE. Journal of Crystal Growth 98, 209–219.

Albert, M.B., Avery, D., Narin, F., McAllister, P., 1991. Direct validation of citation counts as indicators of industrially important patents. Research Policy 20, 251–259.

Amano, H., Sawaki, N., Akasaki, I., Toyoda, Y., 1986. Metalorganic vapor-phase epitaxial-growth of a high-quality GaN film using an AIN buffer layer. Applied Physics Letters 48, 353.

Amano, H., Kito, M., Hiramatsu, K., Akasaki, I., 1989. p-Type conduction in Mg-doped GaN treated with low-energy electron beam irradiation, LEEBI. Japanese Journal of Applied Physics Part 2 Letters 28 (12), L2112–L2114.

Bader, M.A., 2008. Managing intellectual property in the financial services industry sector: Learning from Swiss Re. Technovation 28 (4), 196–207.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Boyack, K.W., Böner, K., 2003. Indicator-assisted evaluation and funding of research: visualizing the influence of grants on the number and citation counts of research papers. Journal of the American Society for Information Science and Technology 54, 447–461.

Boyack, K.W., Wylie, B.N., Davidson, G.S., 2002. Domain visualization using VxInsight for science and technology management. Journal of the American Society for Information Science and Technology 53, 764–774.

Braam, R.R., Moed, H.F., van Raan, A.F.J., 1991. Mapping of Science by combined co-citation and word analysis. I. Structural aspects. Journal of the American Society for Information Science 42, 233–251.

Buter, R.K., Noyons, E.C.M., van Mackelenbergh, M., Laine, T., 2006. Combining concept maps and bibliometric maps: first explorations. Scientometrics 66, 377–387.

Chen, C., 1999. Visualising semantic spaces and author co-citation networks in digital libraries. Information Processing and Management 35, 401–420.

Chen, C., 2004. Searching for intellectual turning points: progressive knowledge domain visualization. Proceedings of the National Academy of Sciences 101, 5303–5310.

Chen, C., Kuljis, J., Paul, R.J., 2001. Visualizing latent domain knowledge. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews 31, 518–529.

Chen, C., Cribbin, T., Macredie, R., Morar, S., 2002. Visualizing and tracking the growth of competing paradigms: two case studies. Journal of the American Society for Information Science and Technology 53, 678–689.

Davidson, G.S., Hendrickson, B., Johnson, D.K., Meyers, C.E., Wylie, B.N., 1998. Knowledge mining with VxInsight: discovery through interaction. Journal of Intelligent Information Systems 11, 259–285.

de Solla Price, D.J., 1965. Networks of scientific papers. Science 149, 510–515.

Erdös, P., Rényi, A., 1959. On random graphs. Publicationes Mathematicae Debrecen 6, 290–297.

Erdös, P., Rényi, A., 1960. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17–61.

Erdös, P., Rényi, A., 1961. On the strength of connectedness of a random graph. Acta Mathematica Hungarica 12, 261–267.

Fleming, L., Sorenson, O., 2004. Science As a map in technological search. Strategic Management Journal 25, 909–928.

Frantzi, K., Ananiadou, S., Mima, H., 2000. Natural language processing for digital libraries automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries 3, 115–130.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. Sociometry 40, 35–41.

Frietsch, R., Grupp, H., 2006. There's a new man in town: the paradigm shift in optical technology. Technovation 26 (1), 13–29.

Guimera, R., Amaral, L.A.N., 2005. Functional cartography of complex metabolic networks. Nature 433, 895–900.

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

Hashimoto, K., Irie, H., Fujishima, A., 2005. TiO_2 photocatalysis: a historical overview and future prospects. Japanese Journal of Applied Physics 44, 8269–8285.

Hopcroft, J., Khan, O., Kulis, B., Selman, B., 2004. Tracking evolving communities in large linked networks. Proceedings of the National Academy of Sciences 101, 5249–5253.

Institute for Scientific Information (ISI). 〈http://scientific.thomson.com/isi/〉.

Jaffe, A., 1989. Real effects of academic research. American Economic Review 79, 957–970.

Jaffe, A., Trajtenberg, M., 1996. Flows of knowledge from universities and federal labs: modeling the flow of patent citations over time and across institutional and geographic boundaries. Proceedings of the National Academy of Sciences 93 (12), 671–677.

Klavans, R., Boyack, K.W., 2006. Identifying a better measure of relatedness for mapping science. Journal of the American Society for Information Science and Technology 57, 251–263.

Kostoff, R.N., Schaller, R.R., 2001. Science and technology roadmaps. IEEE Transactions on Engineering Management 48, 132–143.

Kostoff, R.N., Eberhart, H.J., Toothman, D.R., 1997. Database tomography for information retrieval. Journal of Information Science 23, 301–311.

Kostoff, R.N., del Río, J.A., Humenik, J.A., García, E.O., Ramírez, A.M., 2001. Citation mining: integrating text mining and bibliometrics for research user profiling. Journal of the American Society for Information Science and Technology 52, 1148–1156.

Leydesdorff, L., Cozzens, S., van den Besselaar, P., 1994. Tracking areas of strategic importance using scientometric mappings. Research Policy 23, 217–229.

Losiewicz, P., Oard, D.W., Kostoff, R.N., 2000. Textual data mining to support science and technology management. Journal of Intelligent Information Systems 15 (2), 99–119.

Mansfield, E., 1972. Contribution of R&D to economic growth in the United States. Science 175, 477–486.

Massini, S., Lewin, A.Y., Greve, H.R., 2005. Innovators and imitators: organizational reference groups and adoption of organizational routines. Research Policy 34, 1550–1569.

Mayer, M., Pareira, T.S., Persson, O., Granstrand, O., 2004. The scientometric world of Keith Pavitt: a tribute to his contributions to research policy and patent analysis. Research Policy 33, 1405–1417.

Milgram, S., 1967. The small world problem. Psychology Today 2, 60–67.

Mima, H., Frantzi, K. Ananiadou, S., 1998. The C-value/example-based approach to the automatic recognition of multi-word terms for cross-language terminology. In: Proceedings of the International Joint Workshop on Cross-Language Issues in AI. Held at 5th Pacific Rim International Conference on Artificial Intelligence, PRICAI' 98. Singapore, 10–21.

Morris, S.A., Yen, G., Wu, Z., Asnake, B., 2003. Time line visualization of research fronts. Journal of the American Society for Information Science and Technology 54, 413–422.

Nakamura, S., Iwasa, N., Senoh, M., Mukai, T., 1992. Hole compensation mechanism of *p*-type GaN films. Japanese Journal of Applied Physics Part 1 31, 1258–1266.

Nakamura, S., Mukai, T., Sengh, M., 1994. Candela-class high-brightness InGan–AlGan double heterostructure blue light-emitting diodes. Applied Physics Letter 64 (13), 1687–1689.

Nakamura, S., Senoh, M., Nagahama, S., Iwasa, N., Yamada, T., Matsushita, T., Kiyoku, H., Sugimoto, Y., 1996. InGaN-based multi-quantum-well-structure laser diodes. Japanese Journal of Applied Physics Part 2 Letters 35 (1B), L74–L76.

Narin, F., Hamilton, K.S., 1996. Bibliometric performance measures. Scientometrics 36, 293–310.

Newman, M.E.J., 2004. Fast algorithm for detecting community structure in networks. Physical Review E 69, 066133.

Niosi, J., 1999. Fourth-generation R&D: from linear models to flexible innovation. Journal of Business Research 45, 111–117.

Peters, H.P.F., van Raan, A.F.J., 1993a. Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. Research Policy 22, 23–45.

Peters, H.P.F., van Raan, A.F.J., 1993b. Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. Research Policy 22, 47–71.

Porter, A.L., 2005. QTIP: quick technology intelligence processes. Technological Forecasting & Social Change 72, 1070–1081.

Rosenberg, N., 1974. Science, invention, and economic growth. Economic Journal 84, 90–108.

Skupin, A., 2004. The world of geography: visualizing a knowledge domain with cartographic means. Proceedings of the National Academy of Sciences 101, 5274–5278.

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.

Small, H.G., Griffith, B.C., 1974. The structure of scientific literatures: I. Identifying and graphing specialties. Science Studies 4, 17–40.

Sorenson, O., Fleming, L., 2004. Science and the diffusion of knowledge. Research Policy 33, 1615–1634.

Storto, C., 2006. A method based on patent analysis for the investigation of technological innovation strategies: the European medical prostheses industry. Technovation 26 (8), 932–942.

Sveikauskas, L., 1981. Technological inputs and multifactor productivity growth. Review of Economics and Statistics 63, 275–282.

Tijssen, R.J.W., 2002. Science dependence of technologies: evidence from inventions and their inventors. Research Policy 31, 509–526.

Tryk, D.A., Fujishima, A., Honda, K., 2000. Recent topics in photoelectrochemistry: achievements and future prospects. Electrochimica Acta 45, 2363–2376.

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., van Leeuwen, T.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.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of "small-world" networks. Nature 393, 440–442.

White, H.D., Lin, X., Buzydlowski, J.W., Chen, C., 2004. User-controlled mapping of significant literatures. Proceedings of the National Academy of Sciences 101, 5297–5302.

Williams, R., Edge, D., 1996. The social shaping of technology. Research Policy 25, 865–899.

Zhou, P., Leydesdorff, L., 2006. The emergence of China as a leading nation in science. Research Policy 35, 83–104.