2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies.

Detecting Research Fronts Using Different Types of Weighted Citation Networks

Katsuhide Fujita[1], Yuya Kajikawa[1], Junichiro Mori[1], Ichiro Sakata[1,2]
[1]School of Engineering, The University of Tokyo, Tokyo, Japan
[2]Policy Alternatives Research Institutes, The University of Tokyo, Tokyo, Japan

Abstract--**In this paper, we investigate the performance of types of weighted citation network for detecting emerging research fronts by a comparative study. Some types of citation network, such as direct citation, co-citation and bibliographic citation were tested in some research domains like complex networks. In this paper, some types of citation networks were constructed for each research domain, and the papers in those domains were divided into clusters to detect the research front. Additionally, we employ some measures for evaluating the research fronts to weighted citation networks. For instance, average publication years and similarities of keywords are effective measures to detect research fronts. By introducing these measures as weights of citation networks to the citation network, we can detect research fronts and promising fields compared with the non-weighted citation networks. We perform a comparative study to investigate the performance of type of weighted citation networks for detecting emerging research field. Especially, we evaluate the performance of each type of weighted citation networks in detecting a research front by using the following measures of papers in the cluster: visibility, measured by normalized cluster size, speed, topological relevance, and density.**

I. INTRODUCTION

Recently, the number of academic papers increases exponentially [7], and each academic area becomes specialized and segmented. Davidson, Hendrickson, et al. [6] show this situations as follows: "For most of history, mankind has suffered from a short age of information. Now, in just the infancy of the electronic age, we have begun to suffer from information excess." Therefore, it is hard for researchers to perceive their specialized fields as a whole, and segmentation occurs simultaneously with specialization, which brings a severe problem and also opportunity to find crucial knowledge by integrating different domains. Because the flood of information in the nature of science, there is a strong need for computational tools of science mapping and emerging topic detection. Previous studies have established effective algorithms for creating academic landscapes and for detecting emerging topics for certain research fronts.

Especially, methods of science mapping by citation analysis has been proposed [2,11]. Researchers have also focused on clustering and visualization [4,5,22]. For example, Leydesdorff and colleagues made a large-scale investigation of a set of academic papers [14,15]. Not only creating static academic landscapes, topological and semantic analysis of a citation network also helps us to focus on significant movements in research fronts and emerging research fields in a broad context [20].

The other approach is to detect emerging clusters of densely connected papers. De Solla Price employed the concept of a research front, a research domain under development where papers cite each other densely [7]. Scientists tend to cite the most recently published articles in their paper, therefore, the network belonging in research fronts on recent work becomes very tight. In a given field, a research front refers to the body of articles that scientists actively cite. Researchers have been studying quantitative methods that can be used to identify and track a research front as it evolves over time. Small and Griffith showed that activated scientific specialists generate clusters of co-cited papers [25]. Braam et al. also investigated the topics discussed in co-cited clusters by analyzing the frequency of indexing terms and classification codes occurring in these publications [3].

On the other hand, citation patterns between papers give some effects to detect emerging research domains. By Shibata et al. [21], a comparative study was performed to investigate the performance of methods for detecting emerging research fronts between three types of citation network, co-citation, bibliographic coupling, and direct citation. Three types of citation networks were constructed for each research domain, and the papers in those domains were divided into clusters to detect the research front. Direct citation, which could detect large and young emerging clusters earlier, shows the best performance in detecting a research front, and co-citation shows the worst. Small proposed a method of tracking and predicting growth areas in the sciences by co-citation analysis that analyzed co-citation networks generated from the top 1% of highly cited papers [23]. Klavans, and Boyack compared the performance of clustering in journal citation networks created by direct citation and co-citation. Their results suggested that a network of direct citation has higher content similarity [12].

However, most of the existing works focus on the no-weighted citation networks, despite weighted citation networks containing some important attributes information of papers have possibilities. The purpose of this paper is to study the characteristics of paper-paper weighted citation networks created by different types of measures as well as their performance in the detection of research fronts. Especially, average publication years, similarities of citation information and similarities of keywords are effective measures for detecting research fronts. By introducing these measures as weights of citation networks to the citation network, we can detect research fronts and promising fields compared with the non-weighted citation networks.

This paper studies the following three research domains. Gallium nitride (GaN) is widely recognized as a recent prominent innovation in the fields of applied physics and material science. Complex network (CNW) analysis is also recognized as pioneering a new research field. Nano-carbon (carbon nanotube [CNT]) is also widely recognized as a recent prominent innovation in the fields of applied physics and material science. We constructed the three types of weighted citation network for each domain and divided the citation networks of each research domain into clusters to detect research fronts. We evaluated the performance of each method in detecting a research front by comparing the visibility, as measured by the normalized cluster size, speed, as measured by average publication year, and topological relevance, as measured by density, of the clusters. By considering the differences, we discuss which type of citation is most suitable for detecting emerging knowledge domains.

The remainder of this paper is organized as follows. First, we describe the overview of research domains. Next, we describe the methodology based on the network clustering and network measures. Then, we present and discuss the performance of the types of weighted citation network for detecting emerging research fronts by a comparative study. Finally, we present our overall conclusions.

II. OVERVIEW OF RESEARCH DOMAINS

Gallium nitride (GaN), Complex network (CNW), and carbon nanotube (CNT) 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. Within a very short period following the mid 1990s, researchers realized applications of GaN as blue and green light-emitting diodes (LEDs), ultra violet (UV) and blue laser diodes (LDs) [16-18]. 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. Some papers written by a researcher who worked in a Japanese firm has opened a new route to synthesizing high-quality GaN films having superior optical properties.

The second innovation is CNW, which was recently recognized as a new research field. Previously, CNWs have been researched in the following types of research: graph theory in mathematics, social network analysis in sociology, and applied physics [1,26]. A prominent breakthrough occurred in the last domain, applied physics. Therefore, it can be expected that CNW research in applied physics forms a research front.

The third innovation is CNT, which is useful in nanoscience and nanotechnology, due to superior electrical and mechanical properties. A CNT is a nano-sized carbon molecule having morphology like a tube. Fullerenes are also a well-known nano-sized carbon material having morphology like a ball. The existence of fullerenes was known earlier than that of nanotubes [8]. But after the discovery of the carbon nanotube, the focus of researchers shifted fullerenes to nanotubes. Therefore, if we can detect research fronts that include papers where the discovery of the nanotube is mentioned, we might expect such shift of research focus earlier than competitors. In all of the above cases, earlier detection of research fronts is essential information for both researchers and research and development (R&D) managers to plan their research focus and strategy.

III. METHODOLOGY

TABLE 1: CORE PAPERS THAT OPENED A NEW RESEARCH FRONTIER IN THREE DOMAINS.

Research domain	Core papers
Gallium nitride	(A) NAKAMURA S, 1992, JPN J APPL PHYS PT 1, V31, P1258
Complex networks	(B) Watts DJ, 1998, NATURE, V393, P440
Carbon nanotube	(C) IIJIMA, S, 1991, NATURE, V354, P56

The first step is to collect the data of each knowledge domain and to make citation networks. Citation networks were constructed by direct-citation, co-citation and bibliographic-coupling. After constructing the networks, maximum connected components were extracted from each network. After extracting the maximum components, we divided the papers in the network into clusters. Finally, we evaluated the visibility, defined as normalized size, speed, defined as average publication year, and topological relevance, defined as density, of the clusters to which selected core papers belong. A list of core papers in each domain, which opened a new research frontier, is shown in Table 1.

A. Data Collection

We collected 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 used the Web of Science, which is a Web-based user interface of the ISI's citation databases. We searched the papers using the following terms as queries: *"GaN OR gallium nitride"* for the first domain, *"social networks OR social network OR random networks OR random network OR small-world OR scale-free OR complex networks"* for the second domain, and *"carbon AND (nano* OR micro*)"* for the third domain. In this study, queries were selected according to the following two steps: (a) the representative keyword, such as gallium nitride and social network, is selected and (b) if the definition of its domain is unclear, more keywords, such as random network, small-world, scale-free, and complex networks, were added. Our intention in using so many terms is to retain wide coverage of citation data in order to avoid omission of core papers. 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 citation data. In this paper, queries were selected based on the query expansion [13]. By using many terms data has wide coverage of citation data in order to avoid omission of core papers.

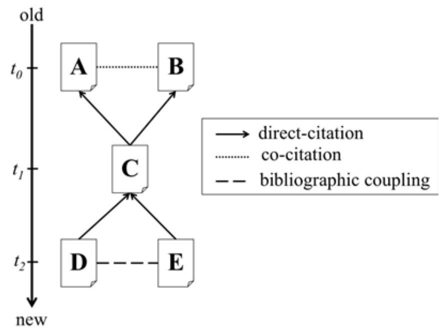

Fig.1: Types of citation.

B. Creating Weighted Citation Networks

We create citation networks by regarding papers as nodes and three types of definitions of citations as edges, as shown in Fig.1. When a paper directly cites anther one as a reference, it calls as the direct citations. In other word, the direct citation is the citing of an earlier paper by a new paper. Co-citation is defined as the edge between two documents cited by the same paper(s) [24]. Bibliographic coupling is defined as the edge between two documents citing the same paper(s) [10]. For example, if both paper *A* and *B* are cited by *C*, there is co-citation between *A* and *B*. And if both *D* and *E* cite *C*, there is bibliographic coupling between *D* and *E* as Fig.1 showing.

We define the citation graphs ***G = (N, E, w)*** comprising a set ***N*** of nodes, which each node N_i representing a paper p_i and a set ***E*** of edges, with each edge E_{ij} directed from the citing node N_i to the cited node N_j. or from the citing node N_j to the cited node N_i. $|E_{ij}|$ means the number of citations between p_i *and* p_j. Usually, the number of direct-citations is one, however, the number of co-citations and bibliographic-couplings is more than one. In other words, we will build the citation networks defined as a weighted non-directed graph, with each paper representing a node and three types of citations representing the edges in the graph. Each node (N_i) has several attributes: paper title, author(s), year of publication (y_i) and journal name, reference information (C_i), and author keywords (K_i).

The network is created in each year enables a time-series analysis of citation networks. When we create citation networks on year *y*, we use the data of papers published from 1960 to *y*, which are available on year *y*. In this paper, only the largest-graph component data is used because this paper focuses on the relationship among papers, and we should therefore eliminate papers that have no link with any other papers.

We also introduced five types of weights to the citation networks; (i) No weight, (ii) Frequency of citations, (iii) Difference of publication years, (iv) Citation similarity, (v) Keyword similarity. The definitions of these weights are defined as follows:

(i) No weight: $w(E_{ij}) = 1$
(ii) Frequency of citations: $w(E_{ij}) = |E_{ij}|$
(iii) Difference of publication years: $w(E_{ij}) = -|y_i - y_j|/10 + 2$ *if* $|y_i - y_j| > 10,\ w(E_{ij}) = 1$
(iv) Reference Similarity: $w(Eij) = Jaccard(C_i, C_j) + 1$
(v) Keyword Similarity: $w(Eij) = Jaccard(K_i, K_j) + 1$
$*\ Jaccard(x, y) = |x \cap y| / |x \cup y|$ (Jaccard similarity is defined by P. Jaccard [9])

By introducing some types of weights based on the attributes, we can detect the research fronts reflected the important attributes, such as new research fronts growing rapidly.

C. Topological Measures in Citation Networks and Network Clustering

In this paper, a fast-modularity clustering proposed by Newman [19] is applied in order to discover tightly knit clusters with a high density of within-cluster edges, which enables the creation of a weighted graph consisting of a large number of nodes. The algorithm is based on the idea of modularity *Q,* which is defined as follows:

$$Q = \sum_s (w_{ss} - a_s^2) = Tr(w) - ||w||^2 \text{ - (1)}$$

where w_{st} is the possibility of the weights of edges in the network that connected nodes in cluster *s* to those in cluster *t*, and $a_s = \sum_t w_{st}$. In the first part of the equation, *Tr(w)*, represents the sum of density of weights of edges within each cluster. A high value of this parameter means that nodes are densely connected within each cluster. The second part of the equation, $||w||^2$, represents the sum of density of weights of edges within each cluster when all edges are placed randomly.

In Newman's method edges that connect clusters sparsely and extract clusters within which nodes are connected densely is cut. A high value of *Q* represents good community division where only dense edges 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 the *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 two clusters is given by

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

In this paper, we stop joining when $\Delta Q < 0$.

D. Topological Measures in Citation Networks

With each citation network, we calculate topological measures such as the number of nodes and edges, maximum of modularity Q_{max}. In addition, for comparing the tendency of some type of weighted citation networks, Visibility (size normalized by the size of the largest component), Speed (average publication year), and Topological Relevance (density) are calculated after clustering to each cluster to which these selected core papers belong. In this paper, we assume that the important front is detected a larger and denser cluster at an earlier stage. When the normalized size of the cluster is larger, we can more easily distinguish the existence of emerging clusters from other clusters. When we have a young average publication year, it means that the cluster can be speedily detected. If the cluster is denser, we can check whether clustering is successful for dividing into clusters.

The size of cluster is defined as normalized to the relative size in order to compare the some types of citation:

$$|N_i \in C| / |N|,$$

where $|N|$ is the total number of entire nodes N and $|N_i \in C|$ is the number of nodes in cluster C.

The density is defined as follows:

$$|E_i \in C| / \binom{|N|}{2} ,$$

where $|E_i \in C|$ is the number of edges, both of the nodes are in cluster C, and $\binom{|N|}{2}$ is the number of combinations from $|N|$ to 2.

IV.RESULTS

A. Basic Topologies of the Networks

Figure 4 shows the time series of Q_{max} of each research domain. In some years, Q_{max} in the weight (iii) is the largest in three types of citations. These results are common regardless of the domain and mean that the weight of the difference of publication years has a "locally dense and globally sparse" structure and can be divided into clusters better than the others. In most of the networks, the Q_{max} becomes smaller as the domain grows. This suggests that the network becomes random as the domain evolves, partly because it becomes denser not only locally but also globally and can't be divided well. Q_{max} becomes higher when extracted clusters do not depend on, in other words, there are many intra-links but fewer inter-links. The low value of Q_{max} means that the network is close to a random network.

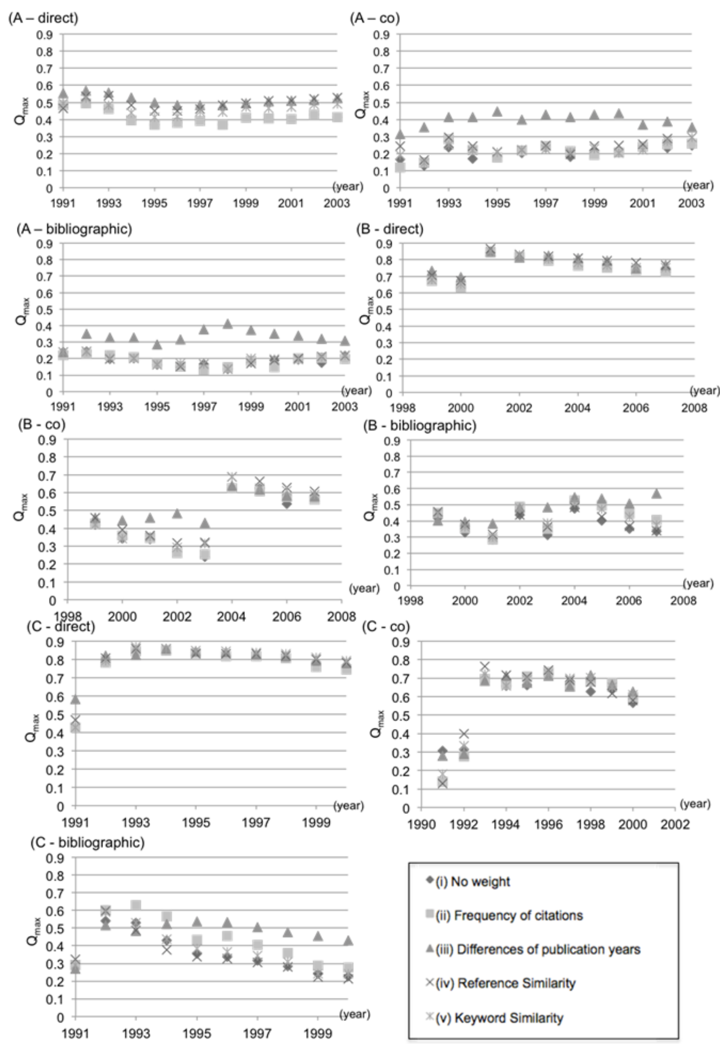

Fig.2: Q_{max} value of each domain: (A) gallium nitride, (B) complex networks, and (C) carbon nanotubes.

TABLE.2: NORMALIZED SIZE, AVERAGE PUBLICATION YEAR, AND DENSITY OF THE CLUSTERS TO WHICH CORE PAPERS BELONG (GAN).

	Direct Citation			Co-Citation			Bibliographic coupling		
Year	Size	Density	Avg. year	Size	Density	Avg. year	Size	Density	Avg. year
(i) No weight									
1992	13	4.167	1989.67	28	12.444	1981.40	37	13.256	1990.00
1993	23	1.391	1990.25	24	7.714	1988.90	39	7.111	1992.50
1994	25	1.027	1991.09	47	4.760	1990.20	43	7.610	1992.78
1995	31	0.519	1992.33	43	5.417	1992.59	43	5.775	1993.92
1996	28	0.260	1994.23	69	2.001	1992.47	40	8.773	1994.71
1997	30	0.168	1994.13	25	1.536	1994.80	39	4.531	1996.27
1998	31	0.130	1994.45	20	0.814	1995.73	41	5.662	1996.77
1999	27	0.103	1995.42	44	1.616	1996.45	9	2.314	1997.78
2000	31	0.072	1996.80	53	0.805	1996.94	38	4.078	1997.21
2001	23	0.068	1998.55	36	1.574	1997.23	6	5.107	1998.26
2002	30	0.053	1999.29	40	1.158	1997.63	29	4.255	1998.81
2003	21	0.050	1999.69	39	0.998	1998.07	28	2.145	1999.10
2004	34	0.036	1999.27	36	1.021	1998.45	27	1.319	2000.35
(ii) Frequency of citations									
1992	13	4.167	1989.67	28	14.333	1991.20	23	11.000	1991.94
1993	23	1.391	1990.25	25	8.186	1992.05	30	12.188	1992.59
1994	25	1.027	1991.09	36	5.448	1992.00	40	7.997	1992.85
1995	31	0.519	1992.33	46	4.879	1993.03	51	7.758	1993.87
1996	28	0.260	1994.23	71	2.920	1994.54	38	5.496	1995.32
1997	30	0.168	1994.13	58	2.537	1995.44	57	5.155	1996.39
1998	31	0.130	1994.45	46	1.396	1996.40	44	4.244	1996.64
1999	27	0.103	1995.42	57	1.932	1997.40	62	7.502	1996.81
2000	31	0.072	1996.80	37	1.076	1998.02	35	2.100	1998.17
2001	23	0.068	1998.55	25	1.596	1997.25	32	6.011	1999.99
2002	30	0.053	1999.29	40	2.681	1998.54	34	1.933	1999.44
2003	21	0.050	1999.69	46	1.442	1999.60	38	4.511	2000.19
2004	34	0.036	1999.27	29	1.783	2000.17	50	3.717	2001.70
(iii) Difference of publication years									
1992	20	2.418	1991.71	11	13.333	1992.00	33	7.036	1991.91
1993	10	2.614	1992.17	33	4.532	1992.10	32	5.226	1992.69
1994	14	1.427	1992.73	25	3.201	1992.30	31	5.479	1994.53
1995	10	0.754	1994.17	20	3.188	1992.16	36	5.886	1994.26
1996	16	0.322	1994.40	29	1.390	1994.71	26	6.160	1994.90
1997	10	0.330	1995.94	22	4.033	1993.61	27	5.517	1994.17
1998	12	0.216	1996.61	37	0.645	1994.19	17	5.536	1994.21
1999	14	0.163	1997.29	21	4.427	1994.26	42	4.342	1995.95
2000	15	0.120	1997.79	18	3.618	1994.79	32	4.259	1995.97
2001	17	0.090	1998.44	25	2.661	1993.59	24	4.358	1995.94
2002	15	0.078	1998.72	21	1.921	1993.03	20	4.234	1996.02
2003	12	0.089	1999.60	19	1.847	1994.17	18	4.211	1996.17
2004	11	0.082	2000.07	30	0.974	1994.66	19	4.259	1997.19

(Continued)

(continued)

	Direct Citation			Co-Citation			Bibliographic coupling		
Year	Size	Density	Avg. year	Size	Density	Avg. year	Size	Density	Avg. year
(iv) Reference similarity									
1992	26	3.091	1991.64	20	14.286	1991.29	23	7.111	1992.00
1993	23	2.251	1992.41	20	9.281	1991.83	30	13.256	1992.50
1994	26	1.161	1993.25	26	5.986	1991.82	40	7.610	1992.78
1995	27	0.683	1994.29	30	6.630	1993.19	51	5.775	1993.92
1996	28	0.387	1995.15	54	2.258	1994.56	38	8.773	1994.71
1997	26	0.264	1995.85	45	1.770	1995.55	57	4.531	1996.27
1998	28	0.184	1996.85	34	1.733	1996.38	44	5.662	1996.77
1999	29	0.121	1997.70	36	0.522	1997.60	43	3.977	1996.83
2000	20	0.102	1998.27	37	1.735	1996.37	37	4.445	1997.15
2001	28	0.077	1998.62	22	3.351	1997.34	30	5.793	1998.11
2002	31	0.061	1999.37	25	2.612	1996.57	35	10.761	1998.32
2003	20	0.091	1999.54	23	2.312	1997.57	30	3.020	1998.56
2004	21	0.075	1999.74	25	2.112	1995.57	23	4.020	1997.56
(v) Keyword similarity									
1992	23	4.167	1991.67	31	10.545	1991.36	23	7.111	1992.00
1993	18	2.747	1992.36	25	6.753	1991.86	36	10.940	1992.56
1994	19	0.900	1993.05	38	5.047	1992.26	40	7.610	1992.78
1995	30	0.474	1994.11	29	7.625	1993.69	51	5.775	1993.92
1996	23	0.333	1995.21	67	2.033	1994.49	38	8.861	1994.89
1997	24	0.265	1995.75	34	3.035	1995.03	57	4.531	1996.27
1998	24	0.130	1996.89	40	1.983	1996.26	44	5.662	1996.77
1999	22	0.120	1997.33	27	2.919	1997.23	18	2.870	1996.55
2000	24	0.086	1997.95	30	2.713	1995.99	34	4.869	1997.09
2001	23	0.094	1998.48	25	1.649	1998.52	32	5.107	1998.26
2002	15	0.063	1998.62	24	0.857	1999.24	31	5.246	1997.80
2003	23	0.064	1999.68	26	1.057	1998.24	30	4.160	1999.30
2004	21	0.065	1999.71	34	1.142	1998.30	50	1.319	2000.35

TABLE.3: NORMALIZED SIZE, AVERAGE PUBLICATION YEAR, AND DENSITY OF THE CLUSTERS TO WHICH CORE PAPERS BELONG (COMPLEX NETWORK).

	Direct Citation			Co-Citation			Bibliographic coupling		
Year	Size	Density	Avg. year	Size	Density	Avg. year	Size	Density	Avg. year
(i) No weight									
2000	24	1.310	1999.57	0	0.000	0.00	0	0.000	0.00
2001	12	0.392	2000.37	0	0.000	0.00	0	0.000	0.00
2002	12	0.325	2001.26	0	0.000	0.00	19	10.650	2001.42
2003	12	0.139	2001.90	0	0.000	0.00	27	5.530	2001.93
2004	17	0.067	2002.74	26	0.385	2001.19	18	8.496	2002.93
2005	16	0.061	2003.60	29	0.458	2002.18	34	2.750	2003.34
2006	17	0.049	2004.40	33	0.429	2002.88	33	3.340	2004.14
2007	17	0.041	2005.13	26	0.550	2003.95	20	10.422	2005.23
(ii) Frequency of citations									
2000	24	1.310	1999.57	0	0.000	0.00	0	0.000	0.00
2001	12	0.392	2000.37	0	0.000	0.00	0	0.000	0.00
2002	12	0.325	2001.26	0	0.000	0.00	19	10.650	2001.42
2003	12	0.139	2001.90	0	0.000	0.00	27	5.530	2001.93
2004	17	0.067	2002.74	34	0.426	2001.21	18	8.496	2002.93
2005	16	0.061	2003.60	33	0.363	2002.07	34	2.750	2003.34
2006	17	0.049	2004.40	28	0.512	2002.39	33	3.340	2004.14
2007	17	0.041	2005.13	26	0.495	2003.38	20	10.422	2005.23
(iii) Difference of publication years									
2000	17	0.996	1999.45	0	0.000	0.00	0	0.000	0.00
2001	8	0.621	2000.48	0	0.000	0.00	0	0.000	0.00
2002	11	0.391	2001.32	0	0.000	0.00	0	0.000	0.00
2003	9	0.221	2001.98	0	0.000	0.00	11	5.358	2000.60
2004	10	0.131	2002.81	26	0.629	2001.23	14	6.382	2001.47
2005	11	0.092	2003.61	29	0.455	2002.13	13	4.865	2002.11
2006	15	0.055	2004.43	23	0.750	2002.56	8	4.969	2003.61
2007	13	0.053	2005.10	22	0.719	2003.79	10	4.370	2005.91
(iv) Reference similarity									
2000	16	1.333	1999.81	0	0.000	0.00	0	0.000	0.00
2001	8	0.638	2000.47	0	0.000	0.00	0	0.000	0.00
2002	11	0.405	2001.38	0	0.000	0.00	0	0.000	0.00
2003	9	0.220	2002.01	0	0.000	0.00	19	10.784	2001.39
2004	11	0.131	2002.89	21	0.910	2001.79	17	12.154	2002.08
2005	11	0.091	2003.63	23	0.625	2002.48	15	11.423	2002.95
2006	12	0.068	2004.41	14	1.570	2003.37	16	11.369	2003.71
2007	13	0.056	2005.12	16	1.094	2004.11	18	10.904	2004.48
(v) Keyword similarity									
2000	16	1.421	1999.80	0	0.000	0.00	0	0.000	0.00
2001	11	0.465	2000.35	0	0.000	0.00	0	0.000	0.00
2002	12	0.359	2001.28	0	0.000	0.00	0	0.000	0.00
2003	12	0.146	2001.93	0	0.000	0.00	18	11.464	2001.43
2004	15	0.085	2002.84	21	0.854	2001.78	18	11.753	2002.10
2005	15	0.067	2003.62	29	0.458	2002.18	20	7.506	2002.98
2006	16	0.050	2004.39	24	0.725	2003.25	17	10.619	2003.72
2007	15	0.048	2005.11	25	0.588	2004.03	18	10.573	2004.48

TABLE.4: NORMALIZED SIZE, AVERAGE PUBLICATION YEAR, AND DENSITY OF THE CLUSTERS TO WHICH CORE PAPERS BELONG (CARBON NANOTUBE).

	Direct Citation			Co-Citation			Bibliographic coupling		
Year	Size	Density	Avg. year	Size	Density	Avg. year	Size	Density	Avg. year
(i) No weight									
1992	14	1.333	1991.87	30	4.053	1991.55	17	12.253	1991.87
1993	10	0.796	1992.69	14	4.486	1992.40	10	4.453	1992.34
1994	11	0.294	1993.29	29	0.752	1992.31	6	2.230	1992.94
1995	10	0.266	1993.99	36	0.523	1992.69	4	2.145	1993.32
1996	10	0.197	1994.57	14	0.376	1992.79	3	1.977	1993.72
1997	10	0.152	1995.28	14	1.631	1994.41	2	1.988	1994.06
1998	11	0.117	1996.13	22	0.388	1993.93	39	0.138	1995.63
1999	13	0.082	1996.94	15	1.419	1996.02	41	0.119	1996.35
2000	13	0.071	1997.80	23	0.691	1996.40	63	0.068	1996.85
(ii) Frequency of citations									
1992	14	1.333	1991.87	30	4.053	1991.55	15	12.952	1991.81
1993	10	0.796	1992.69	15	4.048	1992.38	12	4.981	1992.39
1994	11	0.294	1993.29	16	2.046	1992.69	6	3.244	1992.95
1995	10	0.266	1993.99	15	2.137	1993.11	26	3.644	1993.91
1996	10	0.197	1994.57	21	1.904	1993.42	2	2.269	1993.70
1997	10	0.152	1995.28	23	1.402	1993.30	2	1.726	1994.10
1998	11	0.117	1996.13	22	0.644	1994.65	2	1.847	1994.63
1999	13	0.082	1996.94	17	1.151	1995.87	50	1.101	1996.49
2000	12	0.082	1997.14	23	0.658	1996.36	72	0.164	1996.95
(iii) Difference of publication years									
1992	12	1.477	1991.85	30	4.053	1991.55	18	12.391	1991.88
1993	10	0.787	1992.69	11	3.249	1991.74	13	4.232	1992.25
1994	8	0.440	1993.46	10	0.938	1992.06	9	6.167	1992.71
1995	9	0.387	1994.04	19	1.451	1992.89	5	6.098	1992.75
1996	9	0.211	1994.61	17	1.209	1993.57	5	5.927	1992.63
1997	10	0.259	1995.33	23	0.723	1994.00	7	5.013	1993.24
1998	11	0.123	1995.20	14	1.328	1995.15	9	6.417	1993.80
1999	12	0.190	1996.00	17	0.863	1995.03	9	6.863	1994.41
2000	12	0.176	1996.93	13	1.524	1996.17	8	6.134	1994.52
(iv) Reference similarity									
1992	9	1.947	1991.95	30	4.053	1991.55	23	6.280	1991.87
1993	10	0.831	1992.71	12	5.475	1992.50	11	4.559	1992.33
1994	8	0.446	1993.49	14	2.618	1992.75	17	9.412	1993.38
1995	10	0.239	1993.88	13	2.781	1993.21	15	10.313	1993.97
1996	10	0.186	1994.49	13	1.998	1993.76	14	10.289	1994.60
1997	9	0.170	1995.36	13	1.933	1994.43	14	10.242	1995.30
1998	10	0.125	1996.22	14	1.384	1995.18	14	9.853	1996.14
1999	12	0.092	1997.09	14	1.557	1996.00	6	3.234	1997.28
2000	13	0.071	1997.90	15	1.340	1996.98	8	0.197	1996.57
(v) Keyword similarity									
1992	9	1.813	1991.95	30	4.053	1991.55	10	15.824	1991.71
1993	10	0.812	1992.70	21	2.656	1992.17	11	4.604	1992.35
1994	11	0.293	1993.29	14	2.467	1992.70	17	9.130	1993.37
1995	10	0.237	1993.90	14	2.407	1993.21	15	10.467	1993.98
1996	11	0.171	1994.45	13	1.862	1993.73	15	10.159	1994.61
1997	11	0.136	1995.16	13	1.836	1994.41	3	1.688	1994.13
1998	11	0.119	1996.17	14	1.296	1995.14	2	1.632	1994.64
1999	12	0.088	1997.06	15	1.416	1996.01	3	0.767	1995.55
2000	13	0.069	1997.89	16	1.142	1996.77	5	0.251	1996.09

B. Performance of Each Method in Detecting Emerging Domains

After clustering the networks, we evaluated the performance of the results in each weighted citation network in detecting emerging research domains. The following domains, to which selected core papers in each domain belong, were tracked: visibility (as normalized size), speed (as average publication year), and topological relevance (as density). The normalized size, average publication year, and density of the clusters to which core papers belong are shown in Table 2.

Direct Citations. In this type of citations, all scores using the weight (ii) are similar to that of the weight (i). This is because that a paper cites anther paper only once. The density and the average year of citation networks using the weight (iii) is better than these of the weight (i) in the early stages of the core paper's publication. However, the normalized size in the citation network using the weight (iii) is smaller a little than that of the weight (i). The normalized size and the density of the weight (iv) are higher a little than those of the weight (i). The normalized size and the density of the weight (v) are also higher a little than those of the weight (i). However, the difference of density between the weight (i) and the weight (v) is smaller than one between the weight (i) and the weight (v).

Co-citations. In this type of citations, the results of comparisons between the weights are almost same as the direct citations. However, there is a time lag in co-citation as pointed out by Hopcroft, Khan, Kulis, & Selman (2004). Therefore, the results of the average years don't show the differences, definitely. On the other hand, the density and the average year of the weight (ii) are better than those of the weight (i). In fact, the number of frequency of occurrences is effective of analyzing the citation networks based on co-citations [24].

Bibliographic Citations. In this type of citations, the results of comparisons between the weights are almost same as the co-citations. The bibliographic coupling could be expected to be best because it could potentially detect more edges earlier than the other two methods. However, the results of bibliographic coupling are slightly worse than direct citation when introducing the weighted citation networks.

V. DISCUSSIONS

A summarize of comparisons of the results is shown in Table 3. The weight (ii) generates a younger average birth year and higher density clusters compared with the weight (i). This means that co-occurrences of citations are effective for generating the larger and dense clusters. In addition, Q_{max} of the weight (ii) is slightly larger than the weight (i). The weight (iii) generates denser clusters than the weight (i), and the average birth year is almost same. Therefore, the weight (iii) generates denser and younger clusters in early stage, and the clusters including core papers don't change as the time passes. The weights (iv) and (v) are almost same tendency compared with the weight (i). The reason of this is that both of the reference similarities and the keyword similarity represent the contents of papers. In addition, the references show the more accurate contents than the author keywords. Therefore, the weight (iv) is slightly better than the weight (v) in the early stages.

VI. CONCLUSIONS

This paper represents a comparative study to investigate the performance of methods for detecting emerging research fronts among weighted citation networks. The weighted citation networks include the frequency of citations, the difference of publication years, the reference and keyword. A case study in three research domains, gallium nitride, complex networks, and carbon nanotubes, was performed. After some types of weighted citation networks were constructed, papers in each research domain were divided into clusters using a topological clustering. We evaluated the visibility, defined as normalized size, speed, defined as average publication year, and topological relevance, defined as density, of the clusters to which selected core papers belong.

By using the weight based on the frequency of citations, young and dense clusters are detected. By using the weight based on the difference of published years, clustering techniques generate denser clusters. By using the weight based on author keywords and reference information are almost same tendency. In addition, the references show the more accurate contents than the author keywords.

TABLE.5: BRIEF RESULT OF COMPARISON OF FIVE TYPES OF WEIGHTS.

	Visibility (normalized size)	Topological relevance (density)	Speed (average birth year)
Direct citation	(iv) > (i) = (ii) > (v) > (iii)	(iv) > (v) > (iii) > (i) = (ii)	(iii) > (i) = (ii) = (iv) = (v)
Co-citation	(iv) > (ii) > (v) > (i) > (iii)	(iv) > (ii) > (v) > (iii) > (i)	(iii) > (ii) > (i) = (iv) = (v)
Bibliographic citation	(iv) > (ii) > (i) > (v) > (iii)	(iii) > (iv) > (v) > (ii) > (i)	(iii) > (ii) > (i) = (iv) = (v)

(Note) (i) No weight, (ii) Frequency of citations, (iii) Difference of publication years,
(iv) Reference Similarity, (v) Keyword Similarity

One of the potential weaknesses of citation analysis to detect emerging research front is a time lag to cite (or be cited). Although, in this article, we analyzed only topological data, semantic similarity analysis based on textual data may have the potential to detect emerging research fronts earlier and more precisely. One of the future work is necessary to compare the performance of a link-based approach, text-based approach, and hybrid approach to detecting emerging research fronts.

REFERENCES

[1] Barabasi, AL., and R. Albert, *"Emergence of scaling in random networks,"* Science. Vol.286, no.5439, pp.509-512, 1999.

[2] Boyack, K.W., R. Klavans, and K. Börner *"Mapping the backbone of science."* Scientometrics, 4(3), pp.351–374, 2005.

[3] Braam, R.R., H.F. Moed, and A.F.J. van Raan *"Mapping of science by combined co-citation and word analysis. i. structural aspects."* Journal of the American Society for Information Science, 42, pp.233–251, 1991.

[4] Chen, C., *"Visualizing semantics paces and author co-citation networks in digital libraries."* Information Processing&Management, 35(2), pp.401–420, 1999.

[5] Chen, C., T. Cribbin, R. Macredie, and S. Morar, *"Visualizing and tracking the growth of competing paradigms: Two case studies."* Journal of the American Society for Information Science and Technology, 53, pp. 678–689, 2003.

[6] Davidson, G.S., B. Hendrickson, D.K. Johnson, C. E. Meyers, and B.N. Wylie, B.N. *"Knowledge mining with VxInsight: Discovery through interaction."* Journal of Intelligent Information Systems, 11, pp. 259–285, 1998.

[7] Derek J. de Solla Price *"Networks of scientific papers."* Science, 149, pp. 510–515, 1965.

[8] Iijima, S., *"Helical microtubules of graphitic carbon,"* Nature 354, pp.56 - 58, 1991.

[9] Jaccard, P., *"The distribution of the flora in the alpine zone."* New Phytologist 11(2):37-50, 1912.

[10] Kessler, M.M., *"Bibliographic coupling between scientific papers."* American Documentation, 14, 10–25, 1963.

[11] Klavans, R., and K.W. Boyack, *"Toward a consensus map of science."* Journal of the American Society for Information Science and Technology. 60(3), pp.455–476, 2009.

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

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

[14] Leydesdorff, L., *"Clusters and maps of science journals based on bi-connected graphs in."* Journal of Documentation, 60(4), pp. 371–427, 2004.

[15] Leydesdorff, L., and I. Rafols, *"A global map of science based on the ISI subject categories."* Journal of the American Society for Information Science and Technology, 60(2), pp.348–362, 2009.

[16] Nakamura, S., *"GaN Growth Using GaN Buffer Layer"*, Japanese Journal of Applied Physics, Vol.30, Issue 10A, pp.1705-1707, 1991.

[17] Nakamura, S., T. Mukai, and M. Senoh, *"Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes,"* Applied Physics Letters, vol. 64, no. 13, pp. 1687-1689, 1994.

[18] Nakamura, S., T. Mukai, M. Senoh, *"Si- and Ge-Doped GaN Films Grown with GaN Buffer Layers,"* Japanese Journal of Applied Physics, Volume 31, Issue 9R, pp. 2883, 1992.

[19] Newman, M.E.J., "Fast algorithm for detecting community structure in networks," PHYSICAL REVIEW, E 69, 066133, 2004..

[20] Shibata, N., Y. Kajikawa, Y. Takeda, and K. Matsushima, *"Detecting emerging research fronts based on topological measures in citation networks of scientific publications."* Technovation, 28(11), pp.758–775, 2008.

[21] Shibata, N., Y. Kajikawa, Y. Takeda, and K. Matsushima *"Comparative study on methods of detecting research fronts using different types of citation."* Journal of the American Society for Information Science and Technology 60(3), pp. 571-580, 2009.

[22] Small, H. *"Visualizing science by citation mapping."* Journal of the American Society for Information Science, 50(9), pp. 799–813, 1999.

[23] Small, H., *"Tracking and predicting growth areas in science."* Scientometrics, 68(3), pp. 595–610, 2006.

[24] Small, H. *"Co-citation in the scientific literature: A new measure of the relationship between two documents."* Journal of the American Society for Information Science, 24, pp.265–269, 1973.

[25] Small, H.G., and B.C. Griffith, *"The structure of scientific literatures: I. identifying and graphing specialties. Science Studies"*, 4, pp.17–40, 1974.

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