Technological Forecasting & Social Change 78 (2011) 274–282

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications

Naoki Shibata [a,*], Yuya Kajikawa [a], Yoshiyuki Takeda [b], Ichiro Sakata [c], Katsumori Matsushima [a]

[a] Innovation Policy Research Center, School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo Ward, Tokyo 113-8656, Japan
[b] Department of Project Management, Faculty of Social Systems Science, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, Chiba Prefecture 275-0016, Japan
[c] Todai Policy Alternatives Research Institute, The University of Tokyo, 7-3-1 Hongo, Bunkyo Ward, Tokyo 113-0033, Japan

ARTICLE INFO

Article history:
Received 29 November 2009
Received in revised form 12 May 2010
Accepted 8 July 2010

Keywords:
Citation analysis
Emerging topic detection
Research front
Regenerative medicine
Embryonic stem cells
Induced pluripotent stem cells

ABSTRACT

In this paper, we detect emerging research fronts in a huge number of academic papers related to regenerative medicine, a field of radically innovative research. We divide citation networks into clusters using the topological clustering method, track the positions of papers in each cluster, and visualize citation networks with characteristic terms for each cluster. Analyzing the clustering results with the average published year and parent–child relationship of each cluster could be helpful in detecting recent trends. In addition, tracking topological measures, within-cluster degree z and participation coefficient P, enables us to determine whether there are emerging knowledge clusters. Our results show the success of our method in detecting emerging research fronts in regenerative medicine, and these results are confirmed as reasonable by experts. Finally, we predict the future core papers, with the potential of many citations, via the betweenness centralities in the citation network of the research into adult and somatic stem cells.

1. Introduction

Stem cell biology has been regarded as promising research for use in regenerative medicine [1]. A stem cell is a special kind of cell that has a unique capacity to renew itself and spawn specialized cell types. Although most cells in our body, such as heart or skin cells, are committed to conduct a specific function, a stem cell is not committed and remains so until it receives a signal to develop into a specialized cell [2]. Their proliferative capacity combined with the specialization ability makes them unique. Researchers have long sought ways to use stem cells to replace damaged or diseased cells and tissues [3]. Research in the stem cell field grew out of findings by Canadian scientists Ernest A. McCulloch and James E. Till in the 1960s [4,5]. Recently, the focus has been on two main types of mammalian stem cells: embryonic stem cells (ES cells) that are isolated from the inner cell mass of blastocysts, and adult or somatic stem cells that are found in adult tissues. The amount of research into stem cells is increasing so rapidly that understanding exactly how far research has progressed has become more difficult.

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 by managing new science and technology [6]. Therefore, for both R&D managers and policy makers, understanding emerging research domains among numerous academic papers has become a significant task. Historically, such tasks have been handled by experts, such as by the so-called Delphi method initiated by the Rand Corporation of the US in the 1950s [7,8]. According to Linstone and Turoff, Delphi is defined as a method for structuring a group communication problem [9]. However, it becomes more difficult to create technological foresight using an expert-based

* Corresponding author.
E-mail addresses: shibata@ipr-ctr.t.u-tokyo.ac.jp (N. Shibata), kaji@ipr-ctr.t.u-tokyo.ac.jp (Y. Kajikawa), yoshiyuki.takeda@it-chiba.ac.jp (Y. Takeda), isakata@ipr-ctr.t.u-tokyo.ac.jp (I. Sakata), matsushima@iijmio-mail.jp (K. Matsushima).

0040-1625/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.techfore.2010.07.006

approach because 1) the amount of academic knowledge is increasing so fast that no expert can capture the entire knowledge structure of a specific knowledge domain; 2) the expert-based approach is expensive and time consuming; 3) the generally accepted definition of a targeted research field is sometimes lacking. Moreover, there have recently been considerable systematic efforts using computer-based approach to capture huge amounts of scientific knowledge. Computer-based methods are not only scalable but also provide us common structures in various research fronts.

In this paper, we develop a computational tool with which to support both R&D managers and policy makers in discovering emerging research fronts among a pile of academic publications. There are two types of computer-based methodology which can complement the expert-based approach: text mining and citation mining. As an example of the text mining, Kostoff et al. analyzed multi-word phrase frequencies and phrase proximities, and extracted the taxonomic structure of energy research [10,11]. In previous works, citation-based approaches were used to describe the network of energy-related journals using journal citation data [12] or journal classification data [13]. Recently, Small explored the possibility of using co-citation clusters over three time periods to track the emergence and growth of research areas, and predict their impending changes [14]. In the citation-based approach, citing and cited papers are assumed to have similar research topics. In this paper, we adopt the citation-based approach.

The citation-based approach is useful to obtain a global overview of research domains [15]. By clustering the citation network, we can detect a research front consisting of a group of papers. Although innumerable review papers give an overview of stem cell research, there still remains room to investigate recent emerging research fronts. The aim of this paper is to detect emerging technologies in the stem cell research domain via citation network analysis. Our results can offer an intellectual basis for constructing a strategy for R&D managers and policy makers.

2. Research methodology

In this section, the proposed methodology of this research is shown, while the analyzing schema is depicted in Fig. 1. Step (1) involves collecting the data of each knowledge domain and step (2) constructs citation networks for each year. The problem as to 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 create a corpus for the research domain. However, it causes a problem, deficiency of relevant papers. It is not always true that a research domain can be represented by a single keyword. To overcome this problem, we use broad queries to retain a wide coverage of citation data: "regenerative medicine*", "ES cell*", "embryonic stem cell*", "embryo-derived stem cell*", "ips cell*", "pluripotent stem cell*", "adult stem cell*" or "somatic stem cell*". As a result, we obtained data from 17,824 papers published before the end of 2008. The number of annual publications is shown in Fig. 2, where black circles, white triangles and white rectangles represent the annual number of all papers, retrieved by the queries. As investigated by previous study [15], most papers dealing similar topics can be retrieved in the citation network even if they are truly radical. A paper which comes from a different discipline but cites traditional and commonly cited previous works can be involved in the retrieved network. It is rare for academic papers to cite no relevant previous papers, and therefore, in most cases, a paper is included in the network. In step (3), only the data of the largest graph component is used, because this paper focuses on the relationship among papers, and we should therefore eliminate papers that have no citation from or to any other papers.

After extracting the largest connected component, in step (4), the network is divided into clusters using the topological clustering method [16], which does not need the heuristic input parameters. Newman's algorithm discovers tightly knit clusters with a high density of links within each cluster, which enables the creation of a non-weighted graph consisting of many nodes. By arranging the citation network into clusters, we can detect a research front consisting of a group of papers. However, in co-citation and bibliographic coupling, core papers are sometimes not included in the largest component, especially immediately after these papers were published [17]. Therefore, we regard direct citations as links in citation networks. After clustering, we visualize the citation networks as in step (5), extract the positions of our paper as in step (6) and detect emerging topics as in step (7). In step (5), in order to visualize citation maps, we use a large graph layout (LGL), an algorithm developed by Adai et al. [18], which can be

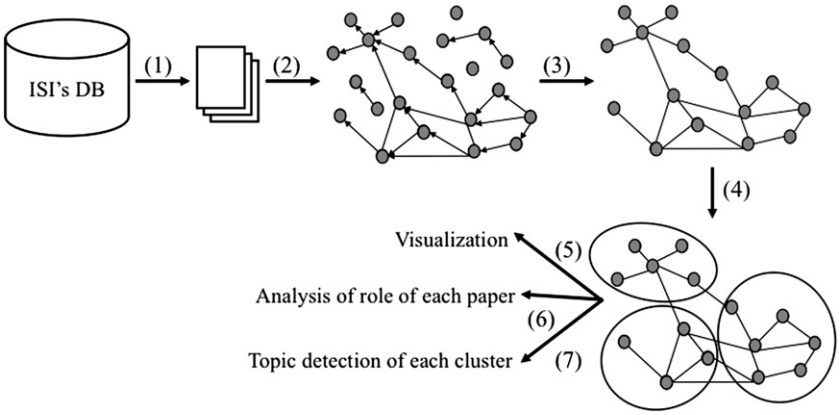

Fig. 1. Methodology proposed in this paper.

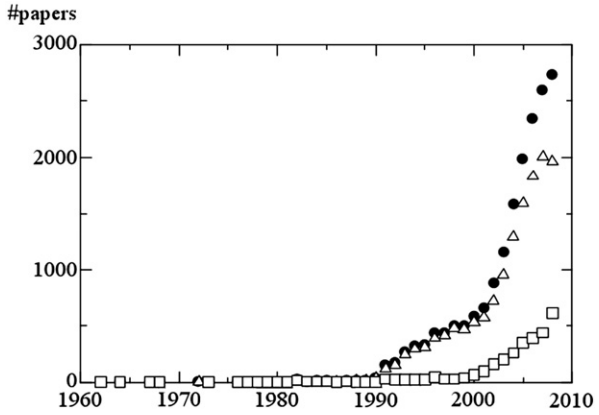

Fig. 2. Number of annual papers including "solar cell" in the title or abstract.

used to dynamically visualize large networks comprised of hundreds of thousands of nodes and millions of links. This algorithm also 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 intra-cluster links in the same color in order that clusters can be intuitively understood.

In step (6), 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 [19]. This method is based on the idea that nodes with identical roles should be in similar topological positions. These two properties can easily be calculated after dividing the network into clusters. The within-cluster degree z_i measures how 'well connected' node i is to other nodes in the cluster, and is defined as:

$$z_i \frac{Ki - \overline{K}_{s_i}}{\sigma_{K_{s_i}}} \tag{1}$$

where K_i is the number of links of node i linked to other nodes in its cluster s_i, $\overline{K}_{si}$ is the average of K overall nodes in s_i, and $\sigma_{K_{si}}$ is the standard deviation of K in s_i. z_i is high if the within-cluster degree is high and vice versa. The participation coefficient P_i measures how 'well distributed' the links of node i are among different clusters and is defined as

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

where K_{is} is the number of links from node i to other nodes in cluster s, and k_i is the total degree of node i (the number of links connected to node i). The participation coefficient P_i is close to 1 if its links are uniformly distributed among all the clusters and 0 if all its links are within its own cluster. Guimerà and Amaral applied this analysis to biological networks and heuristically defined seven different universal roles, by different regions in the z–P parameter space. According to the within-cluster degree, they classified nodes with $z \geq 2.5$ as hub nodes and those with $z \geq 2.5$ as non-hub nodes. In addition, non-hub nodes can be naturally divided into four different roles: (A1) ultra-peripheral nodes, namely, those with most of their links within their cluster ($P<0.05$); (A2) peripheral nodes, namely, those with many links within their cluster ($0.05<P\leq 0.62$); (A3) non-hub connector nodes, namely, those with a high proportion of links to other clusters ($0.62<P\leq 0.80$); and (A4) non-hub kinless nodes, namely, those with links homogeneously distributed among all clusters ($P>0.80$). Similarly, hub nodes can be classified into three different roles: (A5) provincial hubs, namely, hub nodes with the vast majority of links within their cluster ($P>0.30$); (A6) connector hubs, namely, those with many links relative to the other clusters ($0.30<P\leq 0.75$); and (A7) kinless hubs, namely, those with links homogeneously distributed among other clusters ($P>0.75$).

In step (7), the method of extracting the characteristic terms for each cluster by Natural Language Processing (NLP), which enables research topic detection, is described. First, candidate terms are extracted by linguistic filtering, using the abstracts of all papers [20]. Linguistic filtering extracts candidate noun phrases, such as *Noun* + *Noun*, (*Adj*|*Noun*) + *Noun*, and ((*Adj*|*Noun*) + |((*Adj*|*Noun*)* (*NounPrep*)?)(*Adj*|*Noun*)*)*Noun*. Subsequently, these noun phrases are weighted by *tf–idf* (term frequency–inverse document frequency), which is often used in information retrieval. The term frequency (*tf*) in the given documents indicates 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 it has a high term frequency in the given document but seldom appears in most documents. The *tf–idf* weight of term i in document j is given by:

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

where $tf_{i,j}$ is the number of occurrences of term i in document j; $idf_i = log\left(N/df_i\right)$ 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 in a certain cluster rather than a certain document, we extend the *tf–idf* weight to clusters, and the *tf–idf* weight of term i in cluster s is given by:

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

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* value in each cluster are regarded as terms characteristic of that cluster.

3. Results

We collected citation data compiled by the Institute for Scientific Information (ISI), which maintains citation databases covering thousands of academic journals and offers bibliographic database services. We looked at both the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI), two of the best sources for citation data. We used the Web of Science, which is a Web-based user interface of ISI's citation databases, which include papers published after 1970. We searched the papers, using the following terms as the query; "regenerative medicine*", "ES cell*", "embryonic stem cell*", "embryo-derived stem cell*", "ips cell*", "pluripotent stem cell*", "adult stem cell*" or "somatic stem cell*". As a result, we obtained data from 17,824 papers published before the end of 2008. The number of annual publications is shown in Fig. 2, where black circles, white triangles and white rectangles represent the annual number of all papers, those relating to "ES cell*", "embryonic stem cell*" or "embryo-derived stem cell*", and those relating to "ips cell*", "pluripotent stem cell*", "adult stem cell*" or "somatic stem cell*" respectively.

After constructing the largest connected component, in step (4), we divided papers into clusters using the topological clustering method. Fig. 3 shows a plot of the sizes and average ages, which indicates the year minus the average publication year, of each cluster and Fig. 4 shows a chronological evolution of each cluster from 1998 to 2008. In Figs. 3 and 4, circles represent clusters, and the size of each circle indicates the relative value of the number of papers in each cluster. In Fig. 4, the percentages from cluster i in year t to cluster j at $t+1$ mean (the number of papers from cluster i at t to j at $(t+1)$)/(the number of papers in cluster i at t). Only clusters where the number of papers ≧ 300 and percent ≧ 30% were shown. For instance, in 2008 there were mainly three large clusters. The papers in the clusters R_4 and R_5 were much more recently published than the other. As shown in Fig. 3, an emerging cluster R_1, which included 1916 papers and was 1.8 years old, appeared in 2004. Fig. 4 showed that R_1 evolved and was divided into R_2 (3045 papers and 2.3 years old) and R_3 (2946 papers and 2.2 years old) in 2007. R_4 (4640 papers and 2.5 years old) and R_5 (2513 papers and 2.4 years old) were the outcome of the evolution of R_2 and R_3 respectively. However, the papers included in the R_4 and R_5 clusters were similar, since they originated from the same cluster, R_1. In other words, these results suggested a certain degree of semantic relationship between R_2 and R_3 and between R_4 and R_5.

After clustering, in step (5), citation networks were visualized as shown in Fig. 5. Fig. 5(a) shows the citation network in 2008, where R_4 is colored in yellow and R_5 in sky blue, while Fig. 5(b) shows the chronological evolution of citation networks from 1999 to 2008. In Fig. 5(a), the color and position were calculated based on the entire data through 2008 data and, in order to develop the figure of preceding years, we subtracted the data of for each later year in turn and visualized in the same way. The clusters R_4 (yellow) and R_5 (sky blue) were typically emerged, whereas orange denoted one of the traditional clusters. In the scientific field, there are two types of innovations: namely incremental and radical [21]. The question was which type the solar cell domain was and how could we distinguish it?

In order to distinguish the types of innovation, we tracked the role of hub papers by within-cluster degree z and participation coefficient P. Both z and P of hub papers are large when incremental innovations occur, whereas z is large but P is small in the event of radical innovations [15]. Fig. 6 shows plots of z and P for the top ten papers of the number of 2008 citations in this domain. These

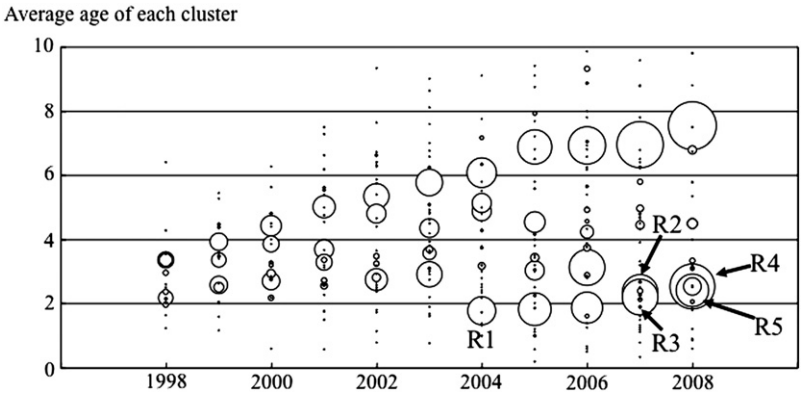

Fig. 3. Cluster size and average age.

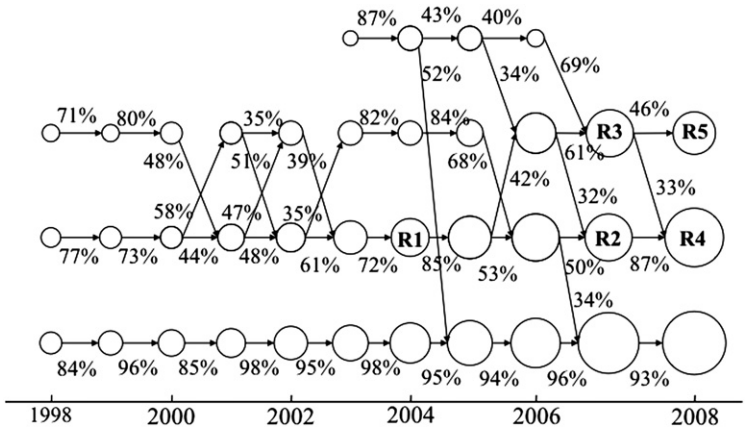

Fig. 4. Chronological visualization of the development of each cluster.

scores vary from year to year. In this case, most hub papers changed position from (A1) ultra-peripheral nodes to (A5) provincial hubs and then to (A6) connector hubs as the domain developed. This domain has had two phases, one of which was radical innovation and the other incremental. In the first phase, which could be assumed by Fig. 3 up through 2004, there were relatively few links among clusters although papers were densely connected within clusters. However in the second phase, after 2004, the growth became incremental and there was a certain amount of links among clusters. These results were also supported by the fact that papers of R_3 in 2007 are divided into R_4 and R_5 in 2008 in Fig. 4.

The next question is how we can identify the emerging clusters by combining the results above and the topics within each cluster. In the final step, we analyze the clustering results and characteristic terms for each cluster. The characteristic terms and hub papers of each cluster in 2004, 2007, and 2008 are shown in Table 1 and the visualization of the clustering result in 2004, 2007, and 2008 is shown in Fig. 7. Only the clusters which satisfied #nodes≧500 for Table 1(a), #nodes≧500 for Table 1(b), and #nodes≧1000 for Table 1(c) were shown. These analyses also supported the results of step (4). An emerging cluster might have all the following features: 1) z of hub papers is large and P is small; 2) hub papers are recently published; and 3) topics, represented by characteristic terms of clusters, differ from other clusters. In 2004, there were mainly three large clusters in Table 1(a) and Fig. 7 (a). Regarding the topics discussed in clusters, the papers in #1 seemed different from other clusters and dealt with applications of embryonic stem cells to human cells. Since clusters dealing with applications to human cells had not appeared before 2004 as densely connected clusters, #2(R_1) was an emerging cluster in 2004.

In 2007, there were mainly two emerging clusters, #2(R_2) and #3(R_3) in Table 1(b) and Fig. 7(b). Topics in cluster #2 (R_2) were embryonic stem cells and those in #3 (R_3) were adult and somatic stem cells. As already mentioned, the embryonic stem cell cluster in 2006 was divided into an adult and embryonic stem cell cluster (#2, R_2) and an adult and somatic stem cell cluster (#3, R_3) in 2007. Since these two clusters had recently published hub papers, and their z values were large and related to different topics, the embryonic stem cell (#2, R_2) and the adult and somatic stem cell clusters (#3, R_3) could be regarded as emerging clusters in 2007.

In 2008, the trends were almost the same as in 2007. There were mainly two emerging clusters, #2(R_4) and #3(R_5) in Table 1(c) and Fig. 7(c). Topics in cluster #2(R_4) were embryonic stem cells, while those in #3(R_5) were adult and somatic stem cells. As already mentioned, most papers in #2(R_4) and #3(R_5) in 2008 were came from #2(R_2) and #3(R_3) in 2007, respectively. In

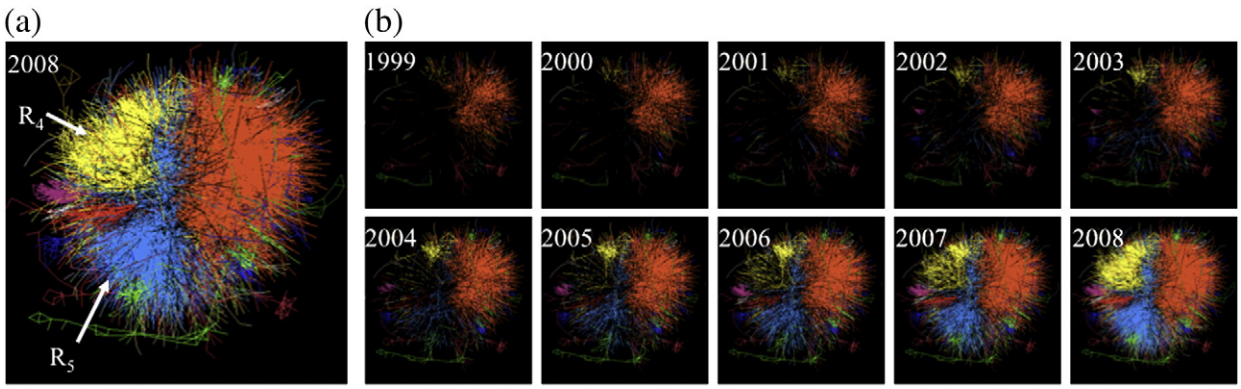

Fig. 5. Visualization of the evolution of the citation network.

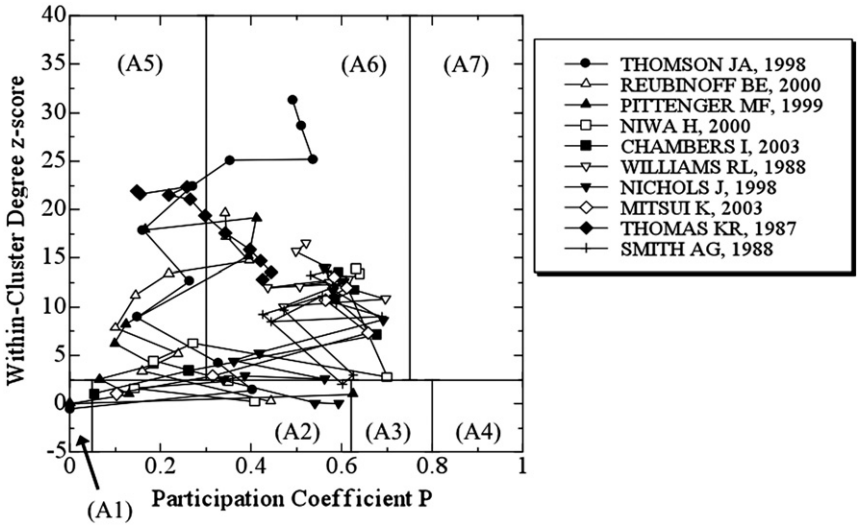

Fig. 6. Changes in the roles of top ten papers in terms of times cited.

particular, cluster #3(R_5) contained papers relating to induced pluripotent stem cells. These findings by Prof. Yamanaka's lab at Kyoto University, "TAKAHASHI K, 2006, CELL, V. 126, P. 663" and "TAKAHASHI K, 2007, CELL, V. 131, P. 861", are considered remarkable.

4. Discussion

As described above, we performed citation network analysis in the regenerative medicine research domain. Our basic idea was that papers dealing with similar topics cite each other and are strongly connected, and those dealing with different topics are

Table 1
The number of papers, average publication year and characteristic terms of major clusters in (a)2004, (b)2007, (c)2008 and (d)2008 #3(R5).

	id	# Papers	Average publication year	Characteristic terms
(a) 2004	1	2093	1997.9	Mice, genes, recombination, cell, mutations, role, expression, functions, beta, development, protein, es cells, stem cells, alpha, allele, wild, dna, loci, levels, cd, animal
	2(R_1)	1916	2002.2	Cell, stem cells, differentiation, es cells, neurons, beta, tissues, expression, oct, culture, transplantation, stem, es, potential, development, genes, mice, source, bodies, marker, disease
	3	924	1999.1	Cell, expression, differentiation, cd, alpha, es cells, beta, role, development, mice, genes, embryos, progenitor, hematopoiesis, stem cells, protein, es, endoderm, receptor, functions, factors
	4	877	1998.9	Cell, methylation, dna, expression, genes, il, es cells, mice, dna methylation, embryos, stat, germ cells, development, factors, culture, gp, es, stem cells, x chromosomes, differentiation, cytokines
(b) 2007	1	4777	2000.1	Mice, cell, genes, expression, es cells, development, role, embryos, recombination, stem cells, mutations, differentiation, beta, es, protein, cd, functions, alpha, stem, levels, dna
	2(R_2)	3045	2004.7	Cell, stem cells, differentiation, neurons, es cells, culture, expression, transplantation, stem, es, development, disease, potential, tissues, protein, mice, beta, genes, lines, application, marker
	3(R_3)	2946	2004.8	Cell, stem cells, oct, genes, differentiation, expression, tissues, cd, beta, insulin, es cells, mice, bone marrow, development, role, dna, protein, methylation, mechanism, studies, progenitor cells
(c) 2008	1	5121	2000.5	Mice, cell, genes, expression, es cells, development, role, cd, beta, differentiation, recombination, stem cells, alpha, es, mutations, embryos, protein, functions, stem, receptor, levels
	2(R_4)	4640	2005.5	Cell, stem cells, differentiation, neurons, tissues, culture, es cells, transplantation, potential, expression, disease, cd, therapies, application, source, stem, development, treatment, marker, bone marrow, mice
	3(R_5)	2513	2005.6	Oct, genes, cell, stem cells, expression, methylation, dna, es cells, dna methylation, role, differentiation, development, protein, mice, germ cells, sox, mechanism, genome, oocytes, self, regulation
(d) 2008 #3(R_5)	1	813	2004.8	Genes, methylation, dna, dna methylation, cell, expression, genome, protein, development, mice, role, x chromosomes, es cells, loci, levels, mechanism, differentiation, modification, stem cells, regulation, histone
	2	751	2005.6	Oct, cell, expression, genes, stem cells, es cells, sox, differentiation, self, germ cells, protein, mice, renewal, development, role, es, stem, factors, promoters, seminoma, marker
	3	685	2006.6	Cell, stem cells, somatic cells, oocytes, embryos, oct, mice, es cells, expression, stem, development, potential, transfer, genes, nucleus, differentiation, state, blastocysts, es, sox, factors

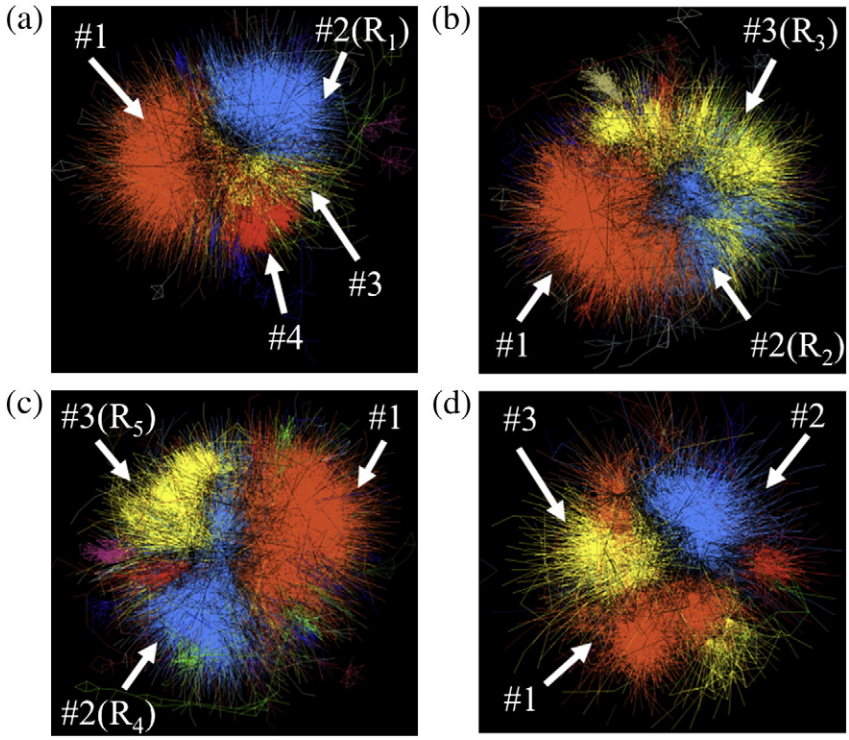

Fig. 7. Visualization of the citation network in (a)2004, (b)2007, (c)2008 and (d)2008 #3(R5).

weakly connected. Therefore, the division of a knowledge domain into strongly connected clusters by citation analysis can help detect emergence. To distinguish innovation types, we tracked the topological positions of hub papers by z and P. In this domain, most hub papers were provincial hubs with many citations within the cluster but few inter cluster. Moreover, we found that most radical innovations occurred before 2004 and most incremental innovations after that, respectively. In order to detect emerging clusters, we extracted as emerging clusters those which had the following features: 1) z of hub papers is large and P is small; 2) hub papers are recently published; and 3) topics, represented by the characteristic terms of clusters, differ from other clusters. As a result, our method could detect the "applications of ES cell to human cells" cluster in 2004 as well as the "ES cells" and "adult and somatic stem cells" clusters in 2007 and 2008 as emerging clusters. These results were reasonable because we used the terms relating to both ES cells and adult and somatic cells. As expected, these results seemed not so valuable. Therefore, we also devised a further examination, dividing the youngest cluster in 2008 (R_5) into sub clusters.

The same methods from steps (4) to (7) were applied to cluster R_5. The characteristic terms and hub papers of each sub-cluster are shown in Table 1(d) and the visualization of the clustering result is shown in Fig. 7(d). In 2008, there were mainly three large clusters in Table 1(d) and Fig. 7(d). Topics in #1 (813 papers, 3.2 years old) concerned general DNA and those in #2 (751 papers, 2.4 years old) concerned Nanog and Oct4, which are the names of genes playing significant roles for pluripotency, while those in #3 (685 papers, 1.4 years old) addressed the issues of induced pluripotent stem cells. After discussion with experts, it was confirmed that our results were helpful for an overview of the research domain. They might be also helpful in deciding what should be further researched and at least could provide new perspectives. Of course, although our method focused only on science publications and did not guarantee that

Table 2
Top five papers in terms of betweenness centrality in 2008.

Paper	Title	Bc in 2008	#Review papers citing
LIU N, 2008, J CELL BIOCHEM, V104, P2348	Identification of genes regulated by nanog which is involved in ES cells pluripotency and early differentiation	0.000317	0
ILIA M, 2003, EXP NEUROL, V181, P159	Expression of the POU domain transcription factor, Oct-6, is attenuated in the adult mouse telencephalon, but increased by neurotoxic damage	0.000280	1
AMBROSI DJ, 2005, J CELL MOL MED, V9, P320	Reprogramming mediated by stem cell fusion	0.000264	2
GOWHER H, 2005, BIOCHEMISTRY-USA, V44, P9899	De novo methylation of nucleosomal DNA by the mammalian Dnmt1 and Dnmt3A DNA methyltransferases	0.000264	2
MA MC, 2005, MAMM GENOME, V16, P391	Analysis of the Xist RNA isoforms suggests two distinctly different forms of regulation	0.000264	0

all emerging research fronts could be commercially, our citation approach could, at least, provide a new viewpoint for constructing roadmaps in domains where the number and speed of publications is high, e.g. in regenerative medicine research.

The final attempt is to predict future core papers in terms of the times cited. A previous study revealed that betweenness centrality correlated with the citations expected in the distant future, especially in an emerging research front [22], e.g. regenerative medicine. Betweenness centrality represents the extent to which a node lies on the paths between other nodes and can also be interpreted as measuring the influence a node has over the spread of information through the network. A paper with a large betweenness centrality bridges unconnected papers, and is therefore anticipated as a previously unexplored seed of innovation. The betweenness centrality of node *i, Bc[i]* is given by following step:

1. To pick up a pair of nodes *s* and *t* other than *i*.
2. To count the number of shortest paths (σ_{st}) between *s* and *t*.
3. To count the number of shortest paths ($\sigma_{st}(i)$) between *s* and *t* through *i*.
4. To calculate the ratio by $\sigma_{st}(i)/\sigma_{st}$.
5. To repeat steps 1 to 4 for all pairs of *s* and *t* and sum up $\sigma_{st}(i)/\sigma_{st}$.

Formally, the betweenness centrality of node *i, Bc*[*i*] is defined as:

$$B_c[i] = \sum_{s \neq i \neq t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node *s* to node *t*, and $\sigma_{st}(i)$ is the number of shortest paths from *s* to *t* traversing *i* [23]. In this paper, we identified the most emerging cluster, R_5, "adult and somatic stem cells," in 2008, constructed a citation network within this cluster, and calculated the betweenness centrality of each paper to predict the candidates for future innovation. The top five papers in terms of betweenness centrality in 2008 are shown in Table 2. According to our prediction, the most promising paper, "LIU N, 2008, J CELL BIOCHEM, V. 104, P. 2348", dealt with the characteristics of Nanog, which is a pluripotent and transcription factor critically involved with the self-renewal of both embryonic and adult/somatic stem cells. According to Dr. Ian Chambers who discovered the mouse Nanog gene, nanog makes stem cells immortal, hence their name from the Tir na nOg legend.

In Table 2, the times cited (the number of citations) from review papers published up through the end of 2008 were described. These five candidates of future innovations had not obtained so many citations by review papers. Our tool could extract seeds of innovations which experts hardly detected. While it does not guarantee that research on that topic will pave the way for new breakthroughs in the research field of regenerative medicine, our hypothesis will be tested in future research in that field.

With this method, we were able to monitor research fronts and detect emerging research by computational calculation alone. Our proposed method demonstrates three advantages over the expert-based approach. The first one is the time and cost to detect. Our approach could detect the same emerging research fronts, such as "applications of ES cell to human cells" cluster in 2004 as well as the "ES cells" and "adult and somatic stem cells" clusters in 2007 and 2008, as described in the previous review reports. These are not surprising but same as experts expected. In other words, the performance of our method was almost same as expert-based approach but much cheaper and faster, if we focused on the first clustering result.

Secondly, computer-based approach is objective, while expert-based approach tends to be subjective. 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. However, computer-based and expert-based approach cannot be comparable by the same perspective. These can be complementary each other. The best way is a hybrid method: obtaining the computer-based results first and discussing by experts based on the objective results.

Finally, our method can extract sub clusters. In this case, the emerging sub clusters could be extracted by dividing cluster #3 (R5) in 2008. As described above, iPS related articles mainly published after 2006 were extracted as an emerging cluster in 2007. As easily imagined, citation networks can be divided as many times as we hope with the clustering algorithm. Experts can zoom in or zoom out based on their objectives in order to discover the seeds of innovations.

Currently, due to specialization and segmentation of research as well as the flood of information, managing R&D activity faces increasing difficulty in understanding a range of diverse research domains and detecting emerging research fronts. However, there is still a lack of researchers, R&D managers, and policymakers who make a point of keeping up with current scientific breakthrough and detecting emerging research fronts. Our topological approach can become a tool for future "Research on Research" (R on R) and can meet the increasing need to discover emerging research fronts in an era of information flooding. Our research can be used as a quantitative method in R on R and technology and innovation management (TIM), and therefore contribute to the research domain.

However, one of the limitations we need to consider is the generality of our results. Our method does not guarantee that extracted emerging articles can attract many other researchers. The outputs are the candidates of seeds of innovation although they can be useful for the decision-making by R&D managers and policymakers.

5. Conclusion

In this paper, we detected emerging research fronts in a huge number of academic papers related to regenerative medicine, which is an area of radically 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 of the

clusters. Analyzing the clustering results with the average published year and parent–child relationship of each cluster can be helpful in detecting emergence. In addition, tracking the two topological measures, within-cluster degree z and participation coefficient P, enabled us to determine whether there were any emerging knowledge clusters. Our results showed our method to be successful in detecting emerging research fronts in regenerative medicine and these results were confirmed as reasonable by experts. Finally, we have predicted potential of future core papers, which have been shown to have many citations, in research involving adult and somatic stem cells by the betweenness centralities in the citation network.

References

[1] T. Reya, S.J. Morrison, M.F. Clarke, I.L. Weissman, Stem cells, cancer, and cancer stem cells, Nature 414 (2001) 105–111.
[2] National Institutes of Health, Stem Cells: Scientific Progress and Future Research Directions, Department of Health and Human Services, June 2001, Retrieved 2/10/09 World Wide Web, http://stemcells.nih.gov/info/scireport/2001report.
[3] National Institutes of Health, Regenerative Medicine, Department of Health and Human Services, August 2006. Retrieved 2/10/09 World Wide Web, http://stemcells.nih.gov/info/scireport/2006report.htm.
[4] A.J. Becke, E.A. McCulloch, J.E. Till, Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells, Nature 197 (1963) 452–454.
[5] L. Siminovitch, E.A. McCulloch, J.E. Till, The distribution of colony-forming cells among spleen colonies, J. Cell. Comp. Physiol. 62 (1963) 327–336.
[6] B.R. Martik, Foresight in science and technology, Technol. Anal. Strategic Manage. 7 (1995) 139–168.
[7] J. Landeta, Current validity of the Delphi method in social sciences, Technol. Forecasting Social Change 73 (2006) 467–482.
[8] A. Kaplan, A.L. Skogstad, M.A. Girshick, The prediction of social and technological events, Public Opin. Q. 14 (1950) 93–110.
[9] H.A. Linstone, M. Turoff, Delphi Method: Techniques and Applications, Addison–Wesley, 1975.
[10] R.N. Kostoff, R. Tshiteya, K.M. Pfeil, J.A. Humenik, G. Karypis, Science and technology text mining: electric power sources, DTIC Technical Report No. ADA421789, in National Technical Information Service, Springfield, VA, 2004.
[11] R.N. Kostoff, R. Tshiteya, K.M. Pfeil, J.A. Humenik, G. Karypis, Power source roadmaps using bibliometrics and database tomography, in Energy 30 (2005) 709–730.
[12] R. Dalpé, F. Anderson, National priorities in academic research-strategic research and contracts in renewable energies, Res. Policy 24 (1995) 563–581.
[13] R.J.W. Tijssen, A quantitative assessment of interdisciplinary structures in science and technology: co-classification analysis of energy research, Res. Policy 21 (1992) 27–44.
[14] H. Small, Tracking and predicting growth areas in science, Scientometrics 68 (2006) 595–610.
[15] N. Shibata, Y. Kajikawa, Y. Takeda, K. Matsushima, Detecting emerging research fronts based on topological measures in citation networks of scientific publications, Technovation 28 (11) (2008) 758–775.
[16] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (2004) 066133.
[17] N. Shibata, Y. Kajikawa, Y. Takeda, K. Matsushima, Comparative study on methods of detecting research fronts using different types of citation, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 571–580.
[18] A.T. Adai, S.V. Date, S. Wieland, E.M. Marcotte, LGL: creating a map of protein function with an algorithm for visualizing very large biological networks, J. Mol. Biol. 340 (1) (2004) 179–190.
[19] R. Guimera, L.A.N. Amaral, Functional cartography of complex metabolic networks, Nature 433 (2005) 895–900.
[20] K. Frantzi, S. Ananiadou, H. Mima, Natural language processing for digital libraries Automatic recognition of multi-word terms: the C-value/NC-value method, Int. J. Digit. Libr. 3 (2000) 115–130.
[21] C.M. Christensen, Innovator's Dilemma: The Revolutionary Book That will Change the Way You Do Business, Harper Collins, New York, NY, 2003.
[22] N. Shibata, Y. Kajikawa, K. Matsushima, Topological analysis of citation networks to discover the future core papers, J. Am. Soc. Inf. Sci. Technol. 58 (6) (2007) 872–882.
[23] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.

Dr. Naoki Shibata is an assistant professor at The University of Tokyo and also a visiting scholar at CSLI, Stanford University. His interests are Management of Technology and Information Science focusing on detecting emerging research fronts by citation analysis. He received his Ph.D degree from The University of Tokyo in 2009. He has an industrial experience as a product manager at CEO office of Rakuten Inc., one of the largest internet companies in Japan. He is a member of The American Society for Information Science and Technology (ASIS&T), Information Processing Society of Japan, and The Japanese Society for Artificial Intelligence.

Dr. Yuya Kajikawa is an assistant professor at the Institute of Engineering Innovation at the School of Engineering in the University of Tokyo. He is also a visiting scholar at research institute for sports industry in Waseda University. He has a PhD in chemical system engineering from the University of Tokyo. He is a multidisciplinary scholar having a wide coverage of knowledge in materials processing, information processing, and technology management. His research background is chemical engineering, and has a number of academic publications in chemical engineering, applied physics, and also materials science. He is also a professional of information science and management science, and has a number of referred journal papers in these disciplines. His current research interest includes R&D management and policy (incl. technology management, technology roadmapping, national innovation system, regional economy), and energy and related technologies (incl. renewable energy, energy policy, thin film processing), and information processing (incl. network analysis, natural language processing, ontology).

Dr. Yoshiyuki Takeda is a research associate at the Department of Project Management at the Faculty of Social Systems Science at the Chiba Institute of Technology. His research interests include Project Management, Regional Innovation System, Web Mining and Information Retrieval. He has a PhD in Computer Science from the Toyohashi University of Technology.

Dr. Ichiro Sakata is a Professor at Todai Policy Alternatives Research Institute, The University of Tokyo. His interests include technology management, technology roadmap and innovation network focusing on rapidly growing sectors including healthcare, solar cell and battery. He received his master's degree form Brandeis University in 1997 and Ph.D from The University of Tokyo in 2003. He has a working experience as a senior policy analyst and policy maker at the Japanese Ministry of Economy, Trade and Industry.

Dr. Katsumori Matsushima is a principal fellow of Innovation Policy Research Center, (IPRC) of the University of Tokyo. He launched his career as an engineer of jet engines of IHI. Dr. Matsushima was engaged in the research on the intellectual production systems at University of Tokyo. He continued his research on CAD/CAM at Technische Universitat Berlin on Alexander von Humboldt scholarship. Later he worked for IBM Japan where he was responsible for the marketing strategy of manufacturing industry, and for personal computer business. In 1997, Dr. Matsushima was appointed as Managing director of PricewaterhouseCoopers Consultant Ltd. Japan, where he has supervised management innovation projects. He joined the University of Tokyo, in 1999 and held position of the director of Institute of Innovation Engineering, and Innovation Policy Research Center. He has been participating in various governmental advisory councils related to Local clusters for economic growth, and technology innovation based on collaboration between universities and industries. Dr. Matsushima is Professor Emeritus of the University of Tokyo.