Detection of Emerging Research Fronts in Solar Cell Research

Yuya Kajikawa[1], Shinji Fujimoto[1], Yoshiyuki Takeda[2], Ichiro Sakata[3] and Katsumori Matsushima[1]

[1] *{ kaji, f_shinji, takeda, sakata}@biz-model.t.u-tokyo.ac.jp, matsushima@iijmio-mail.jp*
Innovation Policy Research Center, Institute of Engineering Innovations, School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8656 (Japan)

[2] *yoshiyuki.takeda@it-chiba.ac.jp*
Department of Project Management, Faculty of Social Systems Science, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, Chiba 275-0016 (Japan)

[3] *isakata@pp.u-tokyo.ac.jp*
Todai Policy Alternative Research Institute, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 (Japan)

Abstract
Science and technology (S&T) roadmaps are an attractive tool in R&D management, and have been widely used during the past decade. S&T roadmaps are typically constructed by gathering and stimulating expert's opinion, but roadmapping is time-consuming and subjective, and therefore computer-based approach is expected to supplement expert-based approach. In this paper, we proposed and demonstrated that the computer-based approach using citation network analysis can be used to depict technology trend, and build the first draft of S&T roadmaps. We perform a case study in solar cell research. We analyzed citation network of energy and solar cell research by clustering the network, visualized the overall structures, extracted emerging research domains there, and track emerging research domains in it by citation network analysis. The possibility and limitation of our approach to roadmapping was discussed. We compared our results by citation network analysis with the existing solar cell roadmap, and showed that citation-based approach can complement expert-based approach.

Introduction
Energy is a key resource to sustain economic activities, our society, and our daily lives. Currently, we have emerging concerns about sustainability and related issues in a number of societal sectors, including the political and economic sectors, universities, and the public. Among them, sustainable production and usage of energy is the prior focus, whose growing interest is partly driven by increasing and fluctuating oil price. Generally speaking, a prominent feature of modern activities is in the creation, dissemination, and application of scientific knowledge, and industrial application of scientific knowledge has contributed to solve social problems by working as seeds of industrial innovations. Issues on sustainable energy is a typical topic where scientific knowledge play a critical role, there is a number of competitive research, and a pile of academic papers and patents have been published. And therefore, academic articles as the outcome of scientific activities and patents to protect intellectual property rights have gained the increasing interest of not only scholars at universities and research institutes but also engineers and policy makers in business and government domains.

However, the rapid pace of science and technology (S&T) growth and globalization has substantially increased the complexity of S&T management. The difficulty has two folds. One is increasing amount of knowledge, which is more than one can handle. That is often called as information flood. Due to this large volume of information, it is becoming hard to understand an overall structure of research and relationships among different researches. It is also hard to detect emerging research domains there. Another difficulty is the increasing speed of publications. Especially, in emerging research domains, we feel a fundamental difficulty in grasping an overall structure and detecting emerging domains, because a situation at a certain point will become old-fashioned and drastically change.

Proceedings of ISSI 2009, pages 71–79.
Edited by B. Larsen and J. Leta.

S&T roadmap is expected to work as a management tool in such situations. S&T roadmap is a represented figure to be constructed for clarifying the direction of research and development (R&D) and sharing future visions on technologies, and promote interdisciplinary collaborations among different participants both in industry and academia. Branscomb and Keller (1998) give the following brief definition of an S&T roadmap, "A consensus articulation of a scientifically informed vision of attractive technology futures." S&T roadmaps are expected to offer a means of communicating visions, attracting resources from business and government, stimulating investigations, and monitoring progress. The prior role of S&T roadmap is to attract resource and research efforts for emerging and promising research domains in an uncertain situation by explicitly showing future research direction. Investment for R&D is an inevitable first step to support scientific activities and to promote technological innovation, and S&T roadmap is expected to work as guidance for those purposes. The role of S&T is especially important because total budget is constrained or has declined (Nemet & Kammen, 2007).

S&T roadmap is often created by an expert-based approach. i.e., a series of hearings, surveys, and workshops, but it is not a rudimentary task to construct a S&T roadmap for making decisions on effective investment in promising and emerging technologies by selecting them from a pile of plausible candidates. Therefore, a computer-based approach is expected to construct S&T roadmap or at least to complement expert-based approach (Kostoff & Schaller, 2001). There is a growing need to analyze a focal research domain by computational tool to assist S&T roadmappin. For that purpose, bibliometric approach such as text mining and citation mining is one of promising tools.

The aim of this paper is to analyze an overall structure of solar cell and to detect emerging research domains there by citation network analysis of academic publications. In previous papers, Kostoff and co-workers analyzed multi-word phrase frequencies and phrase proximities, and extracted the taxonomic structure of energy research (Kostoff, Eberhart, & Toothman, 1999; Kostoff et al., 2002; Kostoff et al., 2005). Other researchers use citation-based approaches to describe the network of energy-related journals using journal citation data (Dalpé & Anderson, 1995) or journal classification data (Tijssen, 1992). These analyses of text mining and citation mining and the results of those by information visualization can depict an overall structure of energy research, but the focus of research is not to detect emerging research domains. On the other hand, Kajikawa and co-workers analyze citation networks and publication trends (Kajikawa et al., 2008; Kajikawa & Takeda, 2008), but overall research structure is still obscure. And it is also worthy to note that these previous papers focus their analysis on academic publications but not patents. Solar cell, which is the target of this paper, is now industrialized and rapidly developing. Therefore, patents are important information source.

In this paper, we visualize the citation network of energy and solar cell researches to depict overall structures of them and emerging research fronts. In the next section, we illustrate brief methodology to analyze citation network academic publications. And then, results on energy research and solar cell research are shown. Finally, we discuss the application potentiality and limitations of our work. We also discuss the relationships of academic papers and patents by comparing key authors publishing core academic papers and those publishing patents.

Research Methodology

Recently, Small (2006) explored the possibility of using co-citation clusters over three time periods to track the emergence and growth of research areas, and predict their near-term

changes. In the citation-based approach, it is assumed that citing and cited papers have similar research topics. We construct networks where nodes are papers and links are citations, and then divide them by clustering .By clustering the citation network, we can detect a research front consisting of a group of papers. However, in co-citation and bibliographic coupling, core papers sometimes were not included in the largest component, especially soon after these papers were published (Shibata et. al, 2009). Therefore, we regard direct citations as links in citation networks.

The first step to perform computer-assisted forecasting is to build relevant corpus. We collected citation data of energy-related academic publications from the Science Citation Index (SCI) compiled by the Institute for Scientific Information (ISI). We used the Web of Science, which is a Web-based user interface of ISI's citation databases. Bibliographic records of patents on solar cell were collected from the web page of European Patent Office (EPO). Our corpus consists of three parts. One is academic publication of entire energy research. Journal Citation Report (JCR) by ISI was used to collect energy-related papers. We included 152,514 papers published in 68 journals categorized as Energy & Fuels in JCR. Another is academic publication of solar cell research. 16,199 records are retrieved and collected by the query of "solar cell*" in SCI. The other is patent of solar cell. We use "solar cell*" as the query and collected 10,749 patents.

After obtaining the above data, the citation networks of academic publications were converted into a non-weighted, non-directed network. Finally, the network is divided into clusters using the topological clustering method (Newman & Girvan, 2004; Newman & Girvan, 2004). The clustering is not fuzzy. A good partition of a network into clusters means there are many within-cluster links and minimal between-cluster links. After clustering the network, we visualized citation networks by large graph layout (LGL) (Adai et al., 2004). LGL 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, which can be used to visualize large networks in the order of hundreds of thousands of nodes and millions of edges, and We visualize the citation network by expressing inter-cluster links as the same colour. We also analyzed the characteristics of each cluster by the titles and abstracts of papers that are frequently cited by the other papers in the cluster, as well as the journals in which the papers in the cluster were published. We named each cluster and also listed the keywords for each cluster from the titles and abstracts of the top twenty most cited papers in the cluster. The number of papers in each cluster was plotted along the time line to know the technological trend. Average publication paper in the cluster ($year_{ave}$) was also calculated.

Results

Research overview and research fronts in energy research

Before analyzing solar cell research, we analyzed the citation network of energy research to comprehend the position of solar cell research and evaluate it. Figure 1 is the visualized results of energy research. By clustering the network, we obtained main 10 clusters. The largest cluster (cluster #1) is Combustion (Cluster E1). It includes 12,128 papers whose average publication year is around 1996. The reaction mechanism of a flame in turbulent flow is the main topic discussed in Cluster #1. The second largest cluster is Coal (cluster #2), which is the oldest cluster among top 10 clusters. Recent research topics are liquefaction, gasification, coal char, and combustion. The third cluster (cluster #3) is Battery. This cluster is younger the above two clusters. Cluster #4 (Petroleum) is as old as Cluster #2 (Coal). Cluster #5 (Fuel cell) and Cluster #9 (Solar cell) show remarkably young $year_{ave}$. Cluster #6

(Wastewater) seems to be noisy from the perspective of sustainable energy. In this cluster, treatment of wastewater such as textile dye is mainly discussed, while only a small fraction of papers study sustainable energy, e.g., biomass. This inclusion of a noisy cluster is attributed to our selection of corpus, i.e., we simply collected papers from journal categories of ISI to know the global trend of energy research but not form queries. Other clusters are #7 Heat pump, #8 Engine, and #10 Power system. In sum, we can see traditional energy researches in the right hand side of Fig. 1 such as combustion, coal, petroleum, and engine. On the other hand, in the left hand side of the figure, we can see emerging researches to realize renewable energy production and usage such as fuel cell, solar cell, heat pump, and battery. We can evaluate solar cell research as one of the emerging research domains among energy research.

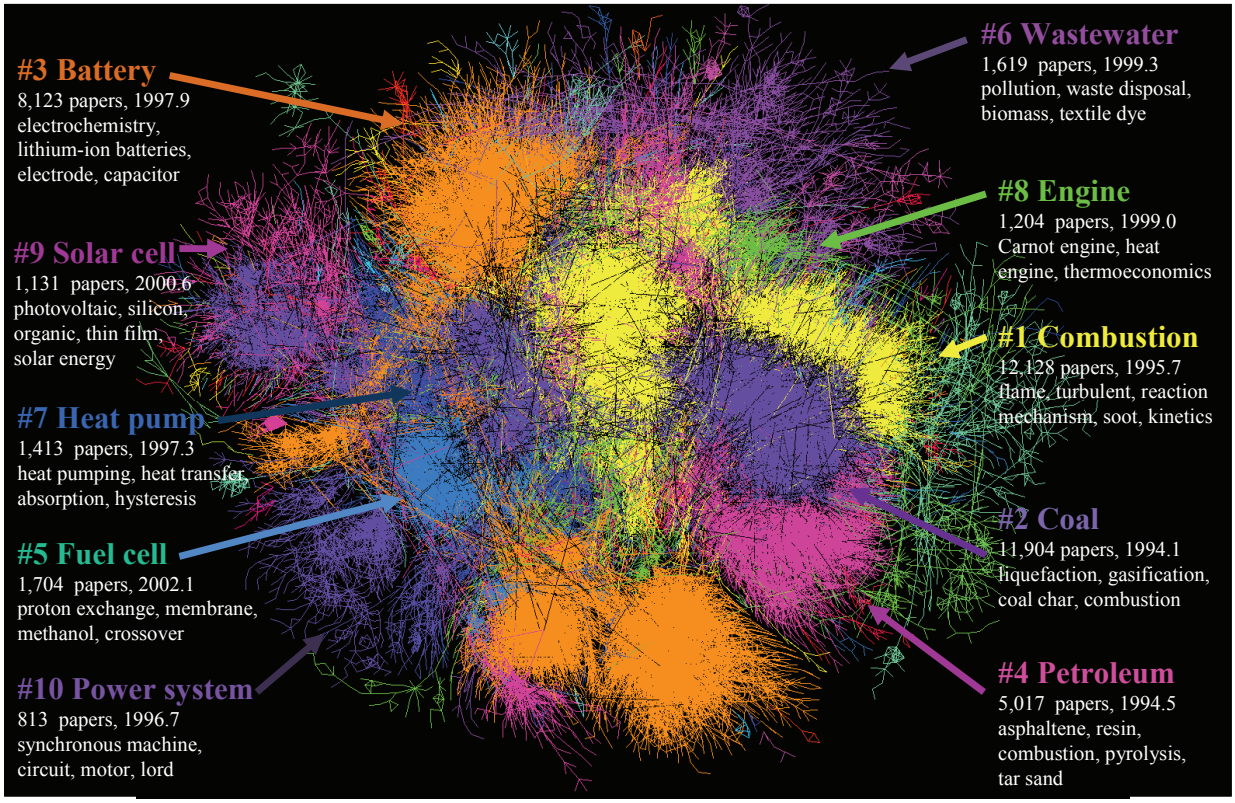

Figure 1. Visualization of citation networks of energy research. Rank of the cluster in the order of the number of papers in that cluster, cluster name, the number of papers in the cluster, average publication years of papers in the cluster, and keywords in the cluster are shown.

Research overview and research fronts in solar cell research

Figure 2 is the visualized results of solar cell research. We can see the main four clusters, #1 Silicon, #2 Compounds, #3 Dye-sensitized, and #4 Organics. The citation network of solar cell research can be divided according to the material used in the cell. In order to analyze the detailed structure of these clusters, clustering is recursively performed for each cluster.

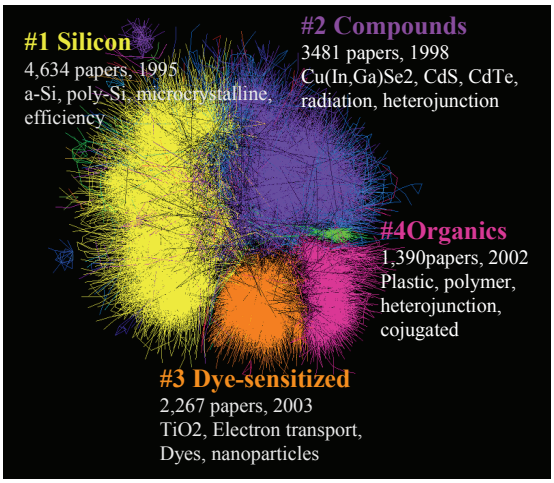

Figure 2. Visualization of citation networks of energy research.

Results of the recursive clustering and its visualization of solar cell research is shown in Fig. 3. By this recursive clustering, we can analyze detailed structure of research in each cluster. For example, Silicon (cluster #1) can be divided into main 5 clusters, i.e., amorphous silicon (a-Si) (cluster #1.1), #1.2 High efficiency cells, #1.3 Modelling, #1.4 Polycrystalline, and #1.5 Limitation and modification of efficiency. Silicon (Cluster #1) is the oldest research clusters among main 4 clusters shown in Figure 2. Its average publication year is 1995, but by using recursive clustering, we can detect emerging research domains even in this traditional domain such as #1.5 Limitation and modification of efficiency where theoretical calculation on the limitation of cell performance and its improvement using new materials structures such as quantum dots and stacking layers are vigorously investigated. In cluster #2 (Compounds), $Cu(In,Ga)Se_2$ (cluster #2.1) and $CuInS_2$ (cluster #2.4) which are called as CIS-type solar cells are emerging ones. In Dye-sensitized cluster (cluster #3), there are a number of emerging sub-clusters. Among them, Electrolyte (cluster #3.2) is especially emerging. In Organics (cluster #4), #4.1 Plastic solar cell and #4.4 Conjugated polymer are emerging and $year_{ave}$ of these clusters are 2004, while it is an old sub-cluster, i.e., #4.3 Cyanine whose $year_{ave}$ is 1997.

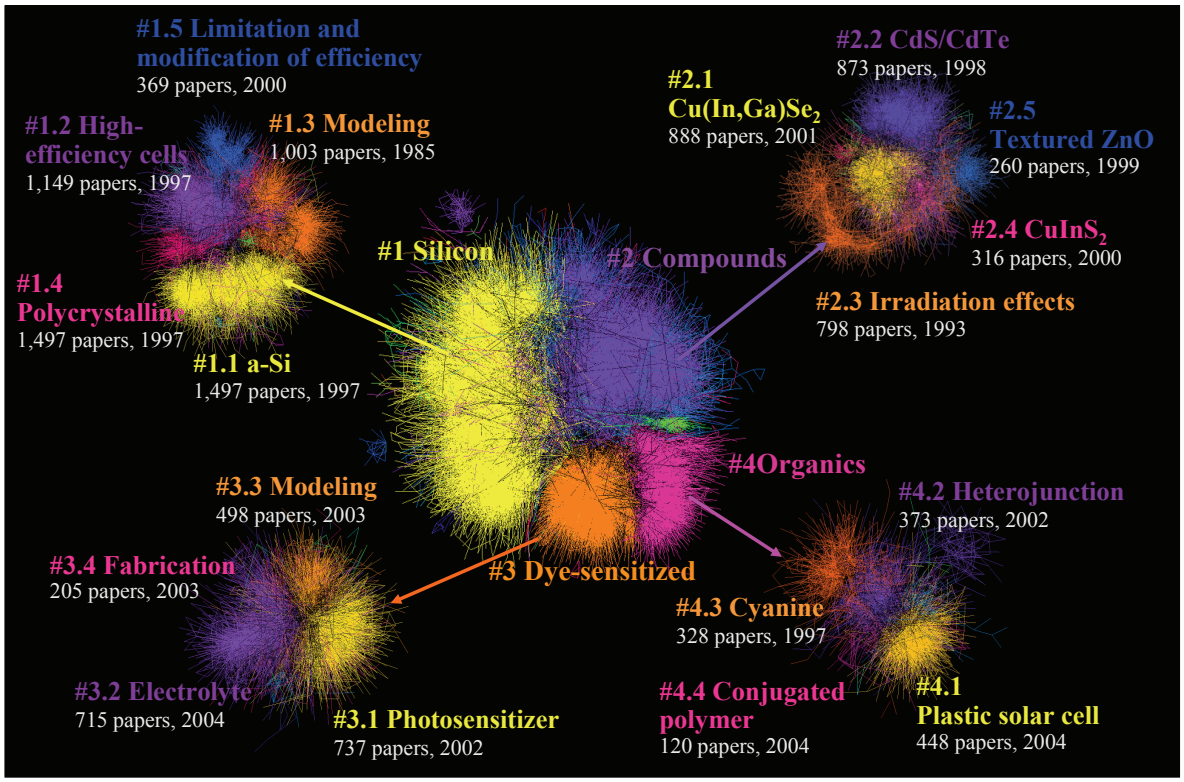

Figure 3. Visualization of subclusters of solar cell research.

Discussion

We compare the above results obtained by citation network analysis with an existing S&T roadmap constructed by an expert-based approach. As for a solar cell roadmap, Photovoltaics (PV) Roadmap toward 2030 (PV2030) was established for a long-term strategy of PV R&D (Aratani, 2005). PV2030 was developed by expert-based approach at the initiative of New Energy and Industrial Technology Development Organization (NEDO) in Japan who is responsible for R&D project planning and formation, project management in Japan. In PV2030, the target is the crystalline Si solar cell, which has the highest market share of PV, and also the high-efficiency GaAs-based or $CuInSe_2$ and $CuInS_2$ (CIS) solar cell. Dye-sensitized cells were the subject of discussion as seed research after 2010. But according to our results, research on electrolyte is especially emerging among a variety of research topics on dye-sensitized solar cell. We also detect the organic solar cell is the emerging domain, which was missed in PV2030. Our analysis successfully can detect the organic solar cell and its sub-domains such as plastic solar cell and conjugated polymer a as an emerging research domains. Therefore, we consider that the computer-based approach can be utilized to offer supplemental information to construct a roadmap using an expert-based approach. The citation network approach is a powerful tool to support experts to construct roadmaps in domains where the number and speed of publications is higher than can be handled such as solar cell research.

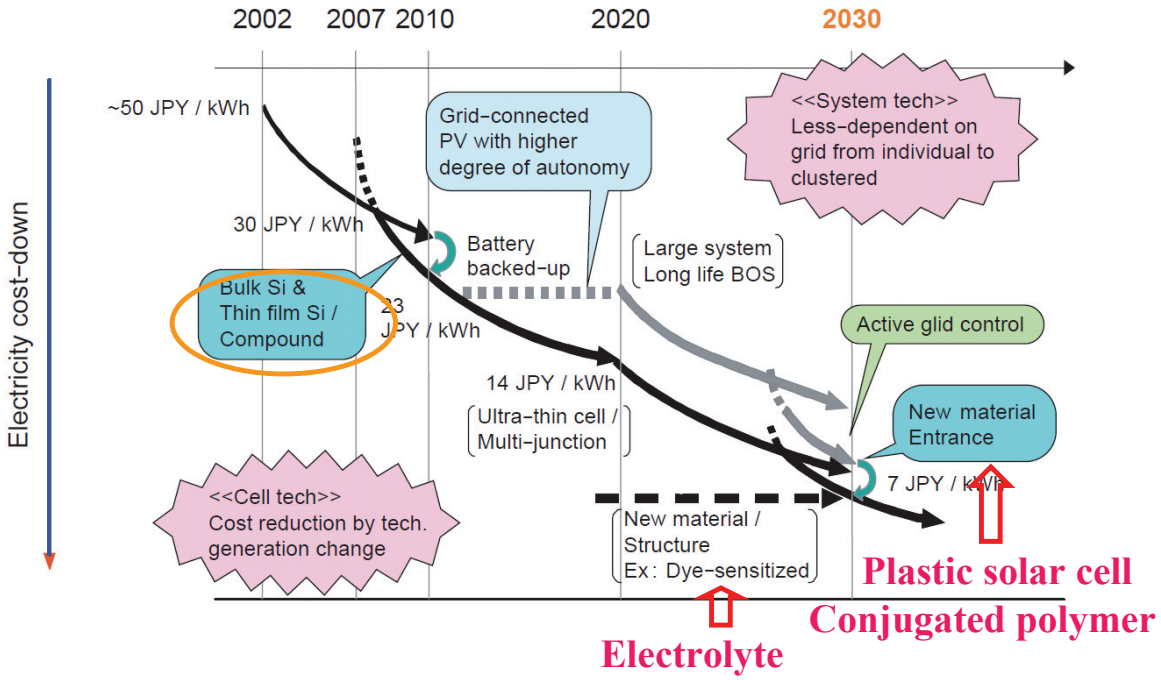

Figure 4. Comparison of PV2030 and our results. Terms written in red are extracted as candidate research topics for the domains shown by the arrows.

There have been a number of researches on citation mining and its visualizations (Boyack, Wylie, & Davidson, 2002; Boyack & Böner, 2003; Chen, 2006). One of plausible application of these bibliometric approaches is management of R&D and S&T policy. For example, Boyack, Wylie, and Davidson (2002) mentioned that "Management of science and technology (S&T) has long been a labour-intensive process, relying extensively on the accumulated knowledge of those within the enterprise. Activities such as technology planning, roadmapping, and the identification of promising or potentially disruptive technologies have been time consuming, and have relied on incomplete information and expert opinion. In addition, the risks associated with poorly managing technology investments have never been greater." However, there is a little effort to compare results obtained by bibliometric approaches with real cases of S&T roadmaps. In this paper, we compared the results obtained by citation network analysis with PV2030 constructed by an expert-based approach, and showed that it can complement the roadmap.

However, we must notice that roadmapping is normative approach and includes backcasting from vision to present status along the pathways to realize the vision (McDowall & Eames, 2006). Citation network analysis is based on existing publications from the past to the present. Therefore, it is a forecasting rather than backcasting. In backcasting, vision is usually given by a top-down approach based on the expert's experience and intuition. For example, we have an option to invest on a traditional research domain for selective budgeting but not on an emerging one, because an emerging research domain is usually highly competitive. But in any case, the feasibility should be tested against existing data and trends. Therefore, we can say that scientifically informed vision included in the Branscomb's definition on roadmap, "a consensus articulation of a scientifically informed vision of attractive technology futures," should be based on the forecasted future of scientific output. In this sense, technological forecasting based on text mining and citation mining can be a key component for S&T roadmapping to construct a reliable roadmap. According to this view, computational tool supports human experts to obtain a scientifically informed vision of attractive technology futures but does not say anything about consensus articulation. While the citation network

approach is a powerful tool to visualize the overall structure of research in a manner that an expert cannot perform and to support experts to construct a roadmap, we should use the results as an intellectual basis for constructing a roadmap but not as a roadmap itself. It is a definitely human task to set a vision and draw a backcasting line for present and future actions. For that purpose, experts and workshops among them are still necessary. But bibliometric approach can assist this process using statistical measures.

One idea is to extract appropriate experts by counting the number of publications. Table 1 is an example for such a purpose. We count the number of academic publications on solar cell research in our corpus. We also show the number of EPO patents by them. We can observe a discrepancy among them, which might reflect diriment skill and knowledge required in academic and industrial world. Therefore, to realize innovation in society based on academic outcome, it might be better to include those different type of researchers. Future study for computer-assisted roadmapping should analyze the relationships between patents and roadmap, patents and academic publications, key inventors of patents and key authors of academic papers.

Table 1. Key inventors publishing patents and academic papers.

Inventor	# of patents	Affiliation	Nationality	# of papers
Barnett Allen M	27	Univ. Delaware.	U.S.A.	26
Martin Green	27	Univ. New South Wales	Australia	204
Wenham Stuart R	24	Univ. New South Wales	Australia	55
Lee Heon	18	Univ Ind & Acad Collaboration	Korea	0
Fath Peter	17	Uni Konstanz	Germany	25
Yang Ru-yuan	14	Southern Taiwan Univ.	Taiwan	2
Frank R	13	Massachusetts Inst. Technology	U.S.A.	3
Kaplow R	12	Massachusetts Inst. Technology	U.S.A.	3
Yi Jun Sin	12	Univ. Sungkyunkwan	Korea	12
Allan Everett Vernie	10	Univ. Australian	Australia	1
William Blakers Andrew	10	Univ. Australian	Australia	43
Luque Lopez Antonio	8	Univ. Madrid Politecnica	Spain	51
Swanson Richard M	8	Univ. Leland Stanford Junior	U.S.A.	18
Kim Dong Hwan	7	Univ. Korea Ind	Korea	4
Park Kyung Hee	7	Univ. Nat. Chonnam Ind. Found	Korea	0
Gu Hal Bon	7	Univ Nat. Chonnam Ind. Found	Korea	1

Conclusion

The aim of this paper is to offer citation network analysis as a technological forecasting methodology and to describe technological trends of solar cell research. We visualized the entire structure of research in energy and solar cell researches, and analyzed emerging research domains there by using citation network analysis. Our analysis confirmed that the solar cell is rapidly growing domains in energy research. By investigating the structure of solar cell research, we observed the existence of main 4 clusters and found that dye-sensitized and organic solar cells are emerging research domains. We also analyzed the detailed structure of them by recursive clustering. At the subcluster level of solar cell research. Among them, electrolyte in dye-sensitized solar cell, and plastic solar cell and conjugated polymer as organic solar cells are especially emerging. By using citation network analysis, we can detect and track emerging research domains among a pile of publications efficiently and effectively. Clustering citation network is useful approach to investigate the detailed structures of a research domain.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 18700240, 2006.

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.

Aratani, F. (2005) The present status and future direction of technology development for photovoltaic power generation in Japan, *Progress in Photovoltaics: Research and Applications*, 13, 463-470.

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.

Branscomb, L. W., & Keller, J. H. (1998). Towards a Research and Innovation Policy, in *Investing in Innovation: Creating a Research and Innovation Policy that Works*, L.W. Branscomb eds., MIT Press.

Chen, C., (2006) CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature, *Journal of the American Society for Information Science and Technology*, 57, 359-377.

Dalpé, R., & Anderson, F. (1995) National priorities in academic research-strategic research and contracts in renewable energies, *Research Policy*, 24, 563-581.

Kajikawa, Y. & Takeda, Y. (2008) Structure of research on biomass and bio-fuels: A citation-based approach, *Technological Forecasting and Social Change* 75, 1349–1359.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K., (2008) Tracking emerging technologies in energy research: toward a roadmap for sustainable energy, *Technological Forecasting and Social Change*, 75, 771-782

Kostoff R. N., Eberhart, H. J., & Toothman, D. R. (1999) Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography, Jo*urnal of American Society for Information Science*, 50, 427-447.

Kostoff, R.N., & Schaller R.R. (2001). Science and Technology Roadmaps*, IEEE Transactions on Engineering Management*, 48, 132-143.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., & Humenik, J. A. (2002) Electrochemical power text mining using bibliometrics and database tomography, *Journal of Power Sources*, 110, 163-176.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., Humenik, J. A., & Karypis, G. (2005) Power source roadmaps using bibliometrics and database tomography, *Energy*, 30, 709-730.

McDowall, W., & M. Eames, M. (2006) Forecasts, scenarios, visions, backcasts and roadmaps to the hydrogen economy: A review of the hydrogen futures literature, *Energy Policy*, 34, 1236–1250.

Nemet, G. F., & Kammen, D. M. (2007) U.S. energy research and development: Declining investment, increasing need, and the feasibility of expansion, *Energy Policy,* 35, 746-755.

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

Newman, M.E.J., & Girvan M., 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation. *Journal of the American Society for Information Science and Technology*, in press.

Small, H. (1996) Tracking and predicting growth areas in science, *Scientometrics*, 68, 595–610.

Tijssen, R. J. W. (1992) A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research, *Research Policy* 21, 27-44 (1992).