Cross-Domain Academic Search using Structural Correspondence Learning

Junichiro Mori
The University of Tokyo
jmori@platinum.u-tokyo.ac.jp
http://academic-landscape.com

1. Project Goal

The vast amount of human knowledge has been accumulated in the form of academic papers. For example, the Web of Science, one of academic citation indexes, currently is indexing tens of millions of academic publications and the number of its records is rapidly growing. With the recent advances in Web technologies, we can easily access the large amount of such scientific data including academic publications, patent, and funding.

Given the large amount of scientific data from a variety of information domains that is easily available from online, one of key questions is how to associate the information from different domains. From the viewpoint of information search, it is a task of associating a source (e.g., paper) from one domain to a target (e.g., patent) from another domain or the other way round.

In this project, we propose a method to automatically associate documents from different domains such as scientific paper, patent, and firm/product of scientific data. The proposed method enables cross-domain academic search on the scientific data. Borrowing ideas from the field of multi-task learning and structural correspondence learning in the field of natural language processing, our approach automatically identifies correspondences among the words from different domains using a small number of so-called concepts. A concept is a commonly used keyword from one domain and another domain, which holds a relevant semantics to respective domains. We also develop a cross-domain academic search engine on top of the proposed method that enable search for related information from different domains of scientific data.

2. Technical breakthrough

Similarity Measure between Documents in the Concept Space

The simple way of associating documents from different domains is to do thesaurus-based mapping among the domains or to model domain-independent concept using comparable corpora. However, vocabulary of one domain (e.g., paper) is often different from vocabulary of another domain (e.g., patent), which cause sparse-overlapping regions of the feature space when mapping documents from two different domains.

One way to overcome this mapping problem is to find a cross-domain representation for documents in different domains, which enables extend the representation of a document by transferring the knowledge between domains. Intuitively, such a cross-domain representation can be considered as a concept space that underlies different domains.

We developed a novel similarity measure that enables compute a semantic similarity between documents from different domains. The similarity measure is based on a mapping function that associates the original representation of a document to the concept space with its cross-domain representation. The mapping function is obtained as follows:

We first collect scientific data including academic papers and patents. We then extract commonly used keywords, so-called *concepts*, across different domains. And we model the correlation between each *concept* and all other words. For this purpose, we train linear classifiers that predict whether or not *concept* occur in a document, based on the other words. A training set is created for each *concept*. The training set contains documents from the domains that *concepts* are extracted. Thus, the classifiers can be considered as cross-domain classifiers. We then reduce the dimension ($|concepts| \times |vocabulary|$) of a parameter matrix

of the linear classifiers to further identify correlations across *concepts*, which gives common substructures among the linear classifiers and can be used as the mapping function.

Academic Landscape System

Using the similarity measure to associate documents from different domains, we have developed the Web-based system called "Academic Landscape System", that gives a user an overview of a research field intuitively by automatically extracting research topics and identifying key researchers and organizations in the research field from hundreds of thousands of academic papers and patents (Fig. 1 and Fig. 2). In addition the system gives a function to associate different domains (e.g., difference research fields or difference sources such as papers and patents). In particular, this function helps a user find the linkage between science and technology or linkage between science and social issues as described in next section.

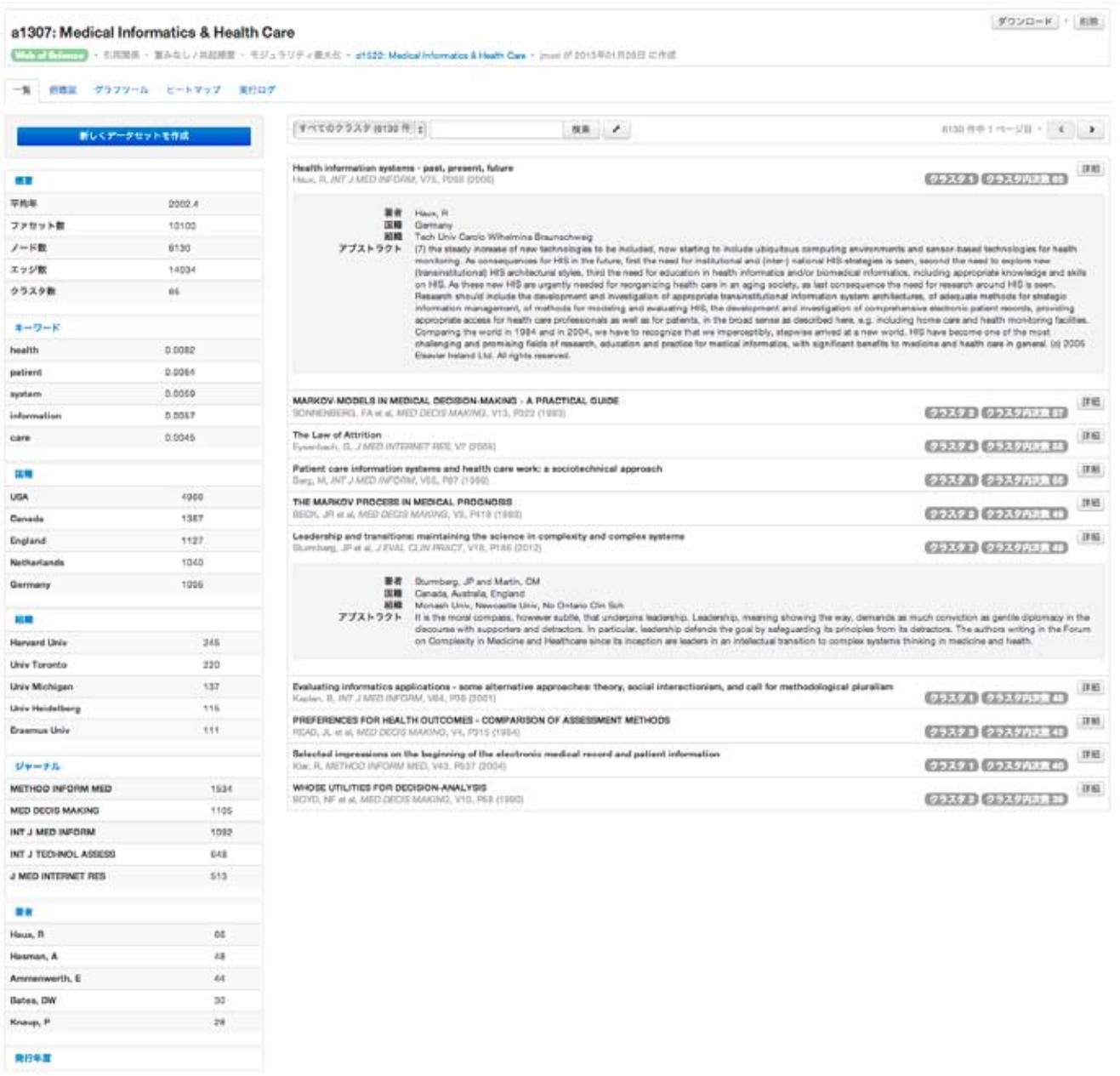

Fig. 1. Academic Landscape System: the system automatically analyzes and associates the large-scale information of scientific papers and patents.

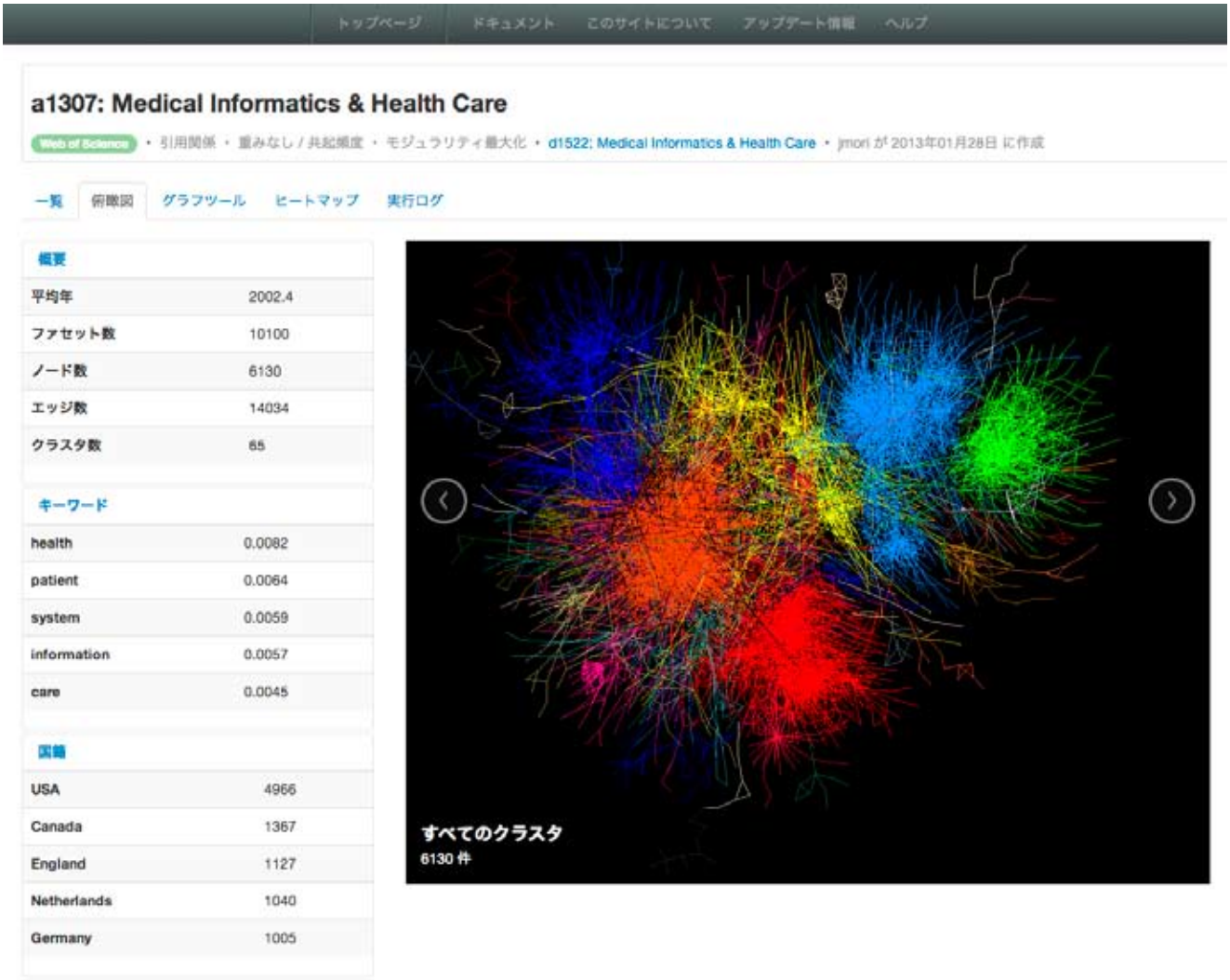

Fig. 2. The Large-scale Citation Network Visualized in the Academic Landscape System

3. Innovative Applications

We describe two of application scenarios of our system in the following.

Linkage between Science and Technology

It is widely accepted that basic research in science provides a fundamental basis for technology-oriented innovations. There are three main layers in the technology-oriented innovation processes; science, technology and industry. Therein, scientists create the seeds of innovation and companies take up these seeds, develop technologies, and then industrialize. For R&D managers and policy makers focusing on future technology, understanding the relationship between science and technology has become a key task in recent years. This understanding helps not only the interaction between R&D and marketing but also the uncertainty reduction, which are signi?cant for technological innovations.

Our system helps a user find the linkage between the scientific outcomes and the pieces of industrial technology by associating papers and patents in the concept space. Fig. 3 shows visualization of several linkages between papers and patents. In vertical axis, each row corresponds to a group of papers in one research field. In horizontal axis, each column corresponds to a group of patents in one technology field. The colored cell, scaling from red (strong) to green (week), indicates similarity between papers and patents. A user effectively identifies the potential linkage between the scienti?c outcomes and the industrial technology.

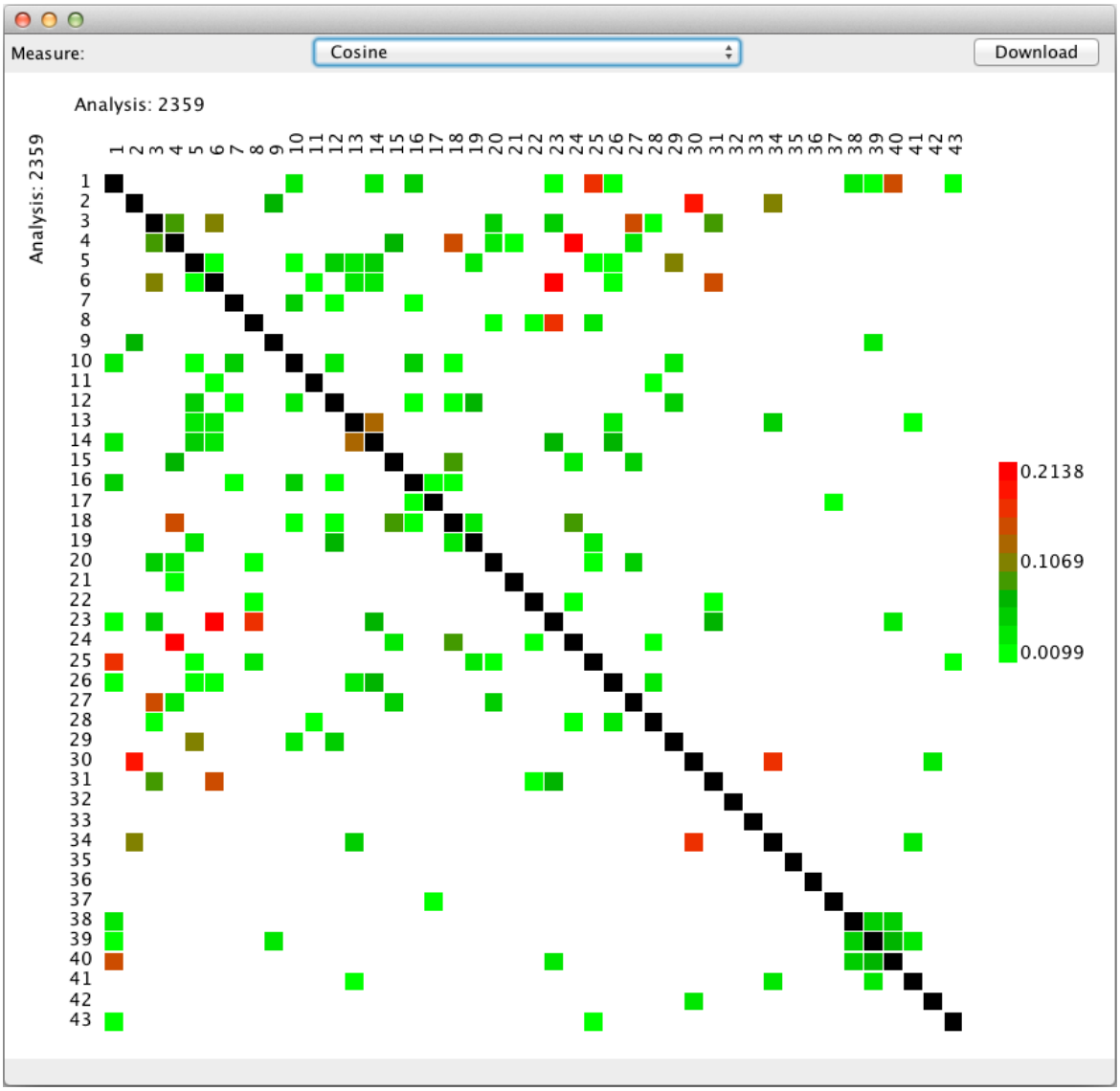

Fig. 3. Visualization of Linkages between Papers and Patents.

Linkage between Science/Technology and Social Issues

The increasingly rapid growth and segmentation of knowledge in the past century have been a problem for both researchers and decision makers. Individual researchers who tend to specialize deeper into very specific fields often experience difficulties in catching up with other research work in their fields of expertise. At the same time, the society's growing concern on social issues and sustainability is putting pressure on decision makers to lead the direction of research and development (R&D) towards key technologies with significant or potential contribution to the future society. To identify these key technologies, decision makers need to better understand the connection between different technologies and social issues. Therefore, there is clearly a strong need for decision support tools to handle the task of finding the linkage.

Our system can provide such managers and policy makers with better understanding of the connections between technology and social issue of interest. Fig. 4 shows identified linkages between robotics (as technology) and gerontology (as social issue) using our system. The result can be further analyzed to identify current key technologies as well as promising research fields and utilized by decision makers in creating technology roadmap, national innovation policy or R&D policy towards future society.

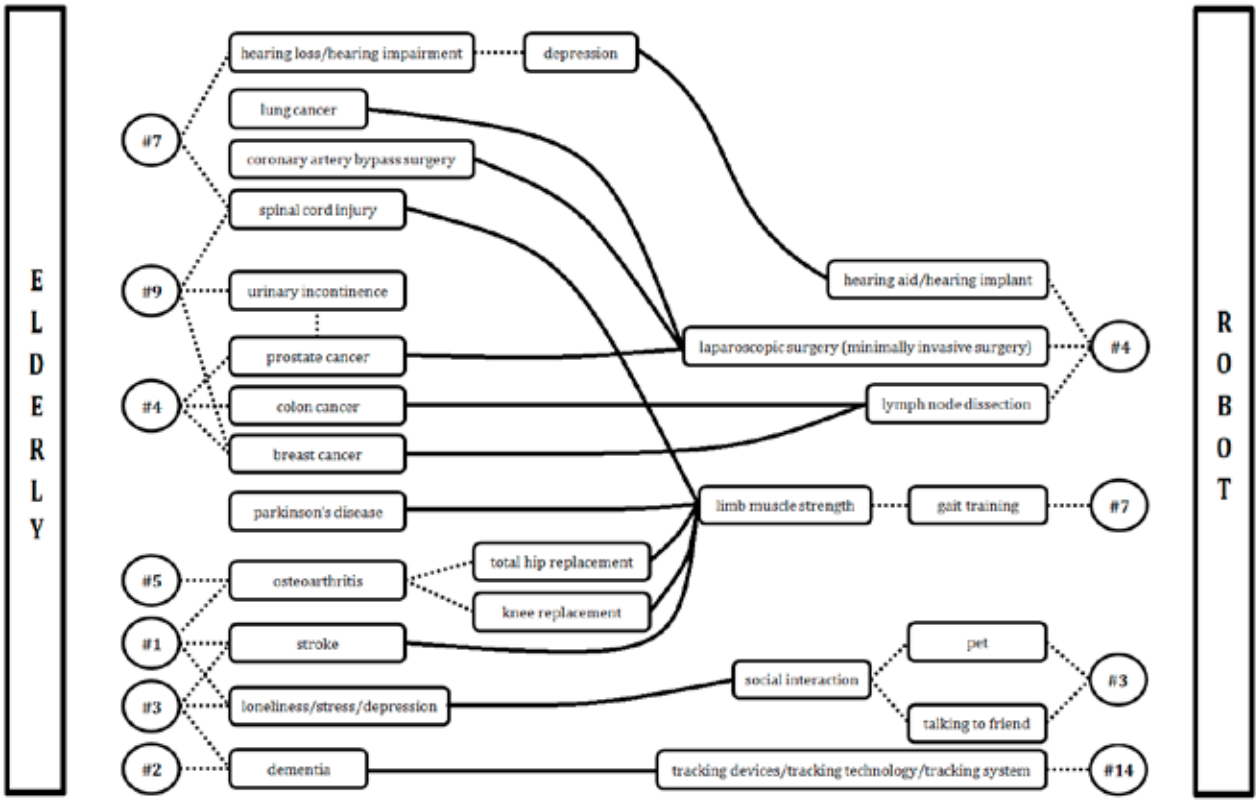

Fig. 4. Overview of linkages between robotics (from 22,519 papers) and gerontology (from 22,864 papers)

4. Academic Achievement

We have published our research results in multiple research domains such as science and information metrics, management science, and technology management. Eight peer-reviewed international conference proceedings have been published. Our paper at the International Conference PICMET 2012 was awarded an outstanding student paper award.

Our Academic Landscape system has been already introduced to several domestic and international institutes such as NEDO (New Energy and Industrial Technology Development Organization), OECD (Organization for Economic Co-operation and Development), IEA (International Energy Agency). It has gained high expectations as a tool for scientific evidence-based planning of policies and corporate strategies. And research and development divisions of several companies including manufacturing industry and ICT industry are currently utilizing the system to decide a strategic planning for their R&Ds.

5. Achievement in Talent Fostering

Several graduate students have been involved in this project. Notably, Vitavin Ittipanuvat, one of our graduate students, who finished his master course during the project was awarded an outstanding student paper award at for his study on "Finding linkage between technology and social issue: A literature based discovery approach" at the International Conference PICMET 2012.

6. Collaboration with Microsoft Research

In our project, we have employed several tools and technologies by Microsoft. We have utilized Microsoft Academic Search APIs and Bing APIs to collect scientific data including the information about academic publications, authors, organizations, and keywords. We believe that the experience gained during the process of developing the proposed method would be useful to further extend the existing Microsoft technologies such as Microsoft Academic Search and Bing.

7. Project Development

The project entitled "Linkage of the Large-scale Science and Technology Information using Structural Correspondence Learning" is on going with the support from the Grant-in-Aid for Scientific Research (24700137) from MEXT.

8. Publications

Paper publication

1) Shino Iwami, Junichiro Mori, Yuya Kajikawa and Ichiro Sakata, "Detection of Next Researches using Time Transition in Fluorescent Proteins", 14th International Conference on Scientometrics and Informetrics (ISSI2013) (to appear)
2) Shino Iwami, Junichiro Mori, Yuya Kajikawa, Tetsutaro Uehara and Ichiro Sakata, "Detection of Promising Fields using Time Transitions in Cryptology", The 22th International Conference for Management of Technology (IAMOT2013)
3) Katuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Detecting Research Fronts using Citation Network Analysis, 2012 Annual Meeting of Institute for Operations Research and Management Sciences (INFORMS2012)
4) Hiroko Nakamura, Yuya Kajikawa, and Shinji Suzuki, Survey and Evaluation of Bibliometrics Methods for Research Planning and Technology Management, 2012 Annual Meeting of Institute for Operations Research and Management Sciences (INFORMS2012)
5) Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Detecting Research Fronts Using Different Types of Combinational Citation, 17th International Conference on Science and Technology Indicators (STI 2012)
6) K. Fujita, Y. Kajikawa, J. Mori, I. Sakata, Detecting Research Fronts Using Different Types of Weighted Citation Networks, In the Proceedings of 2012 Portland International Center for Management of Engineering and Technology (PICMET2012)
7) Vitavin Ittipanuvat, Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Finding linkage between technology and social issue: A literature based discovery approach, Proc. Portland International Center for Management of Engineering and Technology (PICMET2012) (Outstanding Student Paper Award)
8) Vitavin Ittipanuvat, Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, Ichiro Sakata, Measuring relatedness between technology and social issue citation networks, Proc. International Society for Professional Innovation Management (ISPIM2012)