Building a Search Engine for Scientific Projects Survey

Yin, Chengjiu
Research Institute for Information Technology, Kyushu University

Tabata, Yoshiyuki
Research Institute for Information Technology, Kyushu University

Nakatoh, Tetsuya
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

他

出版情報：Proceedings - 2011 IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing, iThings/CPSCom 2011, pp.558-563, 2011-12-01
バージョン：
権利関係：

Building a Search Engine for Scientific Projects Survey

Chengjiu Yin, Yoshiyuki Tabata
Research Institute for Information Technology,
Kyushu University
Fukuoka, Japan
{yin, tabata}@cc.kyushu-u.ac.jp

Xiaobin Wu
Graduate School of Information Science and Electrical Engineering,
Kyushu University
Fukuoka, Japan
2ie10056y@s.kyushu-u.ac.jp

Tetsuya Nakatoh, Sachio Hirokawa
Research Institute for Information Technology, Kyushu University
Fukuoka, Japan
{nakatoh, hirokawa}@cc.kyushu-u.ac.jp

Abstract— **This paper targets the students, who are just beginning to engage in research. With the data-mining technologies, using the data of KAKEN (Grant-in-Aid for Scientific Research of Japan), according to students' learning styles and learners' knowledge levels, we propose to create a "Learning by Searching" search engine to provide suitable knowledge and help students to master research trends. "Learning by Searching" provides newly developed pedagogy to meet the knowledge needs of learners.**

Keywords- Learning by Searching, Search engine, Analysis, Research trends, Data mining

I. INTRODUCTION

In the context of information-seeking, people always consider the search engine. The search engine has been the most significant development in the history of the World Wide Web. When learners face problems in everyday study, they tend to search for answers on the Internet using search engines such as Google or Yahoo for acquiring knowledge through the web. This learning process is called, "Learning by Searching". With the development of the Internet and search engine technology, learning by searching will be a very important learning style. That's where we start this research.

Online searching is becoming a part of our learning processes, and is a necessary skill for students. Although these search engines cater to students' basic knowledge acquisition, they are not categorized into special research area, making it difficult to address the specific, unique needs of each individual learner. The question is how to design better search engines that address users' learning needs and knowledge levels.

An ideal search engine should not only show the retrieval results, but also the analysis. Fortunately, technologies can accelerate learning and boost creativity. With the development of technologies such as data-processing, it is possible to design better search engines to address learning needs. Data-pressing includes functions such as search engine, data mining, recommendations, and image recognition.

For the students, who are just beginning to engage in research, it is very important to do a research survey to collect the information needed, and guide their planning phases of the project. Students can gain knowledge by using the current search engines such as Google or Yahoo. However, these are not search engines which are dedicated to support the acquisition of knowledge according to the special research area.

This study targets the students, who are just beginning to engage in research. By employing several data-mining technologies and the data of KAKEN (Grant-in-Aid for Scientific Research of Japan, http://kaken.nii.ac.jp/ja/searchk.cgi), a "Learn by Searching" approach is proposed and a learning environment with a search engine is developed based on the proposed approach.

This system not only provides a search results, but also analyzes the search results and provides the research trends. Through the retrieval results and its analysis, students can master research trends and decide their research topics. The aim of the system is to analyze large amounts of information in the shortest time, provide and recommend appropriate knowledge, which meets the needs and levels of students.

In this paper, we focus on the KAKEN report, which includes ongoing scientific projects reports, as it suitable to keep up to date with the latest progress. The trend research result gives students insight into the disparate changes occurring in their research field, allowing students to more accurately predict how they should position their own research in the future. It is an important way to help students make decisions about topics of interest.

The organization of the paper is as follows. In the next section, the paper discusses related works and some background. Then the paper describes a Learning by Searching strategy. After that, the paper presents the data-processing of the system. The implementation of the system is introduced afterwards. Finally a conclusion is drawn.

II. RELATED WORKS

Some methods have been proposed to help users to detect the emerging topics in some particular information areas. Bun proposes an Emerging Topic Tracking System (ETTS) which is an information agent for detecting and tracking the emerging topic from a particular information area on the Web [1]. It uses a new TF*PDF (Term Frequency * Proportional Document Frequency) algorithm to detect the changes in the information area of user's interest and generate a summary from the changes back to users from

time to time. This summary of changes will be the latest most discussed issues and it may reveal an emerging topic. Decker et al. uses a semantic approach to propose a method for identifying researchers in the early stages of a research area [2]. It is effective in finding many exact matches of researchers that have major contributions within the research area being identified. The Hierarchical Distributed Dynamic Indexing (HDDI) system mentioned in [3] aims to identify features and methods to improve the automatic detection of emerging trends by generating clusters based on semantic similarity of textual data. The rate of change in the size of clusters and in the frequency and association of features is used as input to machine learning techniques to classify topics as emerging or non-emerging. Collaborative Inquiry-based Multimedia E-Learning (CIMEL) is a multi-media framework for constructive and collaborative inquiry based learning [4]. The semi-automatic trend detection methodology described in [5] has been integrated into the CIMEL system in order to enhance computer science education. A multimedia tutorial has been developed to guide students through the process of emerging trend detection.

Moreover, some researchers use bibliometric methodology to analyze the trends and forecasts in different domains, such as e-commerce, supply chain management and knowledge management [6,7,8]. Using a bibliometric approach, [9] analyzes data mining and CRM research trends from 1989 to 2009 by locating headings “data mining” and “customer relationship management” or “CRM” in topics in the SSCI database. Especially, it uses categories such as publication year, citation, country/territory, document types and the like to explore the differences in the two fields.

It is an important way to help students make decisions about topics of interest, when they begin to do research. In this paper, the KAKEN report is selected as data recourse, we are aiming at the construction of the analysis search engine for research topics. Different from the methods mentioned above, what we can present is not only the emerging topics of the finished researches but also the emerging research topic trends that are occurring. There are 3 features of this system:

1) This system can help learn literature retrieval and analysis of knowledge and methods.
2) This system can help train independent study and build survey literature ability.
3) This system helps students speed up their pace of scientific research and get scientific research achievements early.

III. LEARNING BY SEARCHING

A. *Searching is a natural learning behavior*

Why do people search? A simple answer is that people need to seek for information because they do not have that information in their memory. Searching is a natural learning behavior like listening, speaking, reading or writing. There are many reasons for people to seek for information. Sometime people search for information because of curiosity; that is, they want know why. Sometime people search for information purely for their needs of solving problems or completing tasks. Whatever the actual reason is, the information searching process is a cognitive process that acquires knowledge actively, which is defined as a way of learning, called “learning by searching” in this study.

"Learning by Searching" provides newly developed pedagogy to meet the knowledge needs of learners. There are many kinds of learning strategies, such as learning by attending classes, learning from informal incidents, learning by doing, learning by gaming, learning by searching. Among those learning strategies, learning by searching can foster students the ability of taking the initiative to acquire knowledge. This research advocates learning by searching. It is a method for promoting "active learning" and "discovery learning". Discovery learning is an inquiry-based, constructivist learning theory that takes place in problem solving situations where the learner draws on his or her own past experience and existing knowledge to discover facts and relationships and new truths to be learned [10] Students interact with the world by exploring and manipulating objects, wrestling with questions and controversies, or performing experiments. Bruner suggested that students are more likely to remember concepts if they discover them on their own. This search engine realizes discovery learning and help students learning by themselves. The search engine broadens their sources of knowledge, and improves their self-learning ability. The role of the instructors is changed from givers of information to facilitating student learning.

B. *Categories of knowledge*

Searching can be perceived as a process of acquiring knowledge. When seeking for knowledge on the Internet, people often represent their quests with one of the following “5W1H” questions [11]; that is, “What”, “Where”, Which”, “Who”, “When” and “How”. Researchers have indicated that knowledge can be divided into two categories, one is ability, which includes “know how”, “know what”, “know when” and “know who”; the other is related to knowing where to find knowledge needed, which includes “know where” and “know which”. As shown in Figure 1, ability is being supplemented with knowing where. As knowledge continues to grow and evolve at a speedy pace, access to what is needed is more important than what the learner currently possesses [12].

With the advancement and popularity of the Internet and the search engine technology, online information searching is recognized as being a part of learning and a required learning skill for students [13]. Researchers have attempted to investigate the cognitive processes underlying information searching. For example, State [14] examined the search habits of 72 participants while conducting a total of 426 searching tasks. It was found that search engines were mainly used for checking the learners' own internal knowledge via comparing the knowledge with the searched facts, meaning that the information searching is a learning process rather than simply a way of obtaining information.

This study aims to develop a learning system with a search engine which is able to advice students where to find

the information for solving practical problems. The learning system enables the students to master some of the basic concepts and methods of scientific literature survey during the process of document retrieval. Students can recognize the research trends in depth through accessing the data and viewing the analysis results provided by the learning system.

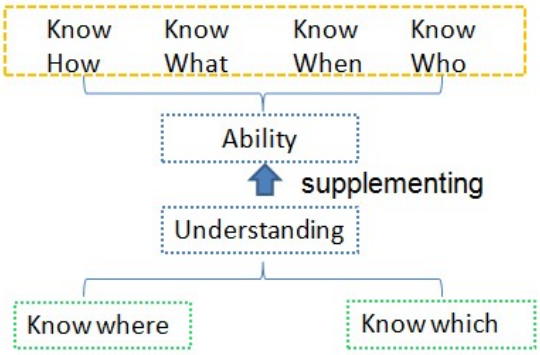

Figure 1. Relation of Knowledge.

IV. DATA-PROCESSING

A. Date Process

1) Indexing: In the present study, an original search engine is constructed using the database of KAKEN, which is available at http://kaken.nii.ac.jp. We downloaded 648,510 documents which range from 1964 to 2011, as of June 24, 2011. Note that the number of documents of 2011 is small, since most of them are not open to public yet. Each document of the project contains, *title, project id, *period,*representative,*members, genre, keywords, outline and amount. We indexed the marked(*) items to construct a search engine. We ignored the representative and members in our implementation. We used special marks such as "y:2009" and "n:SachioHirokawa" to distinguish the year information and the name of the researcher from usual keywords, when they are indexed.

2) Multiple search for matrix display: We explain the interface of our search engine in the next section. The characteristics of our search engine is in the listing of documents obtained as a search result, but in the way that feature words of search result are displayed. User specifies 2 features from "year", "name" or ordinary "word". The number of documents that matched the pair of feature words are shown in a matrix map. Imagine that a list of project outlines are obtained for a query "q", that R1,R2,R3,R4 and R5 are the top 5 researchers and that W1,W2,W3,W4 and W5 are top 5 characteristic words in the search result. The system conducts 5*5 search with "q and Ri and Wj" to calculate the number projects that matches the condition. The number is displayed in the (i,j)-th cell of the matrix.

B. Knowledge Level

According to the students' current ability and knowledge level to determine what kind of retrieval results it is. The knowledge levels are classified into three levels: "Beginner", "Intermediate" and "Advanced". A beginner is someone new to a special field of research. Someone on the Intermediate level knows simple concepts of the field. Someone who is "Advanced" has a wide knowledge about the field. The retrieved results are also classified in to three categories: "Basic Concept", "Middle Concept" (relatively detailed concepts) and "Expert Concept"(detailed concepts). As show in the Figure 2, the system provides a "Basic Concept" for the Beginners, a "Middle Concept" for the "Intermediates" and a "Expert Concept" for the "Advanced ".

Before using this searching system, the students should set-up an account, enter their own profile. Then, based on the users' profile and their interest and past actions, the system will based on their current knowledge level to provide the appropriate knowledge for them.

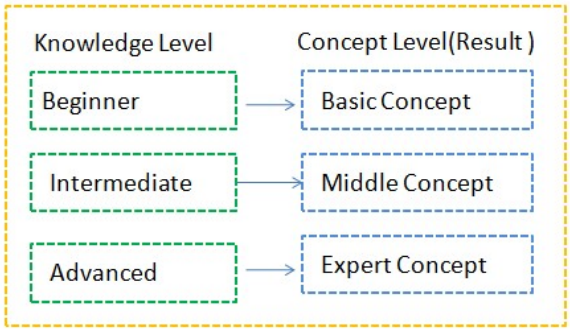

Figure 2. Knowledge and Concept Level.

V. THE IMPLEMENTATION OF THE SYSTEM

A. Interface

As show in the Figure 3, user can control his process by specifying (a) the queries for search and (b) the features to be displayed. The search results are shown in (d) matrix form together with (c) the ordinary listing of documents.

(a) Input query and parameters

1.*Keyword*: the special keyword "z" returns all the documents.
2.*Detail*: display or no-display of parameters can be selected. no-display is default.
3.*Exclusion* of stop words: The frequent words are excluded as stop words.
4. *DB*: choice of data bases.
5.*Axes of matrix*: x-axis and y-axis in matrix display can be chosen from "year", "name", and "word".
6. *Debug*: Displays parameters for debugging.
7.*Sort*: Choose sort function from document frequency or weight(default).
8. *Name*: The number of researchers to be analyzed.
9.*Word*: The number of keywords to be analyzed.

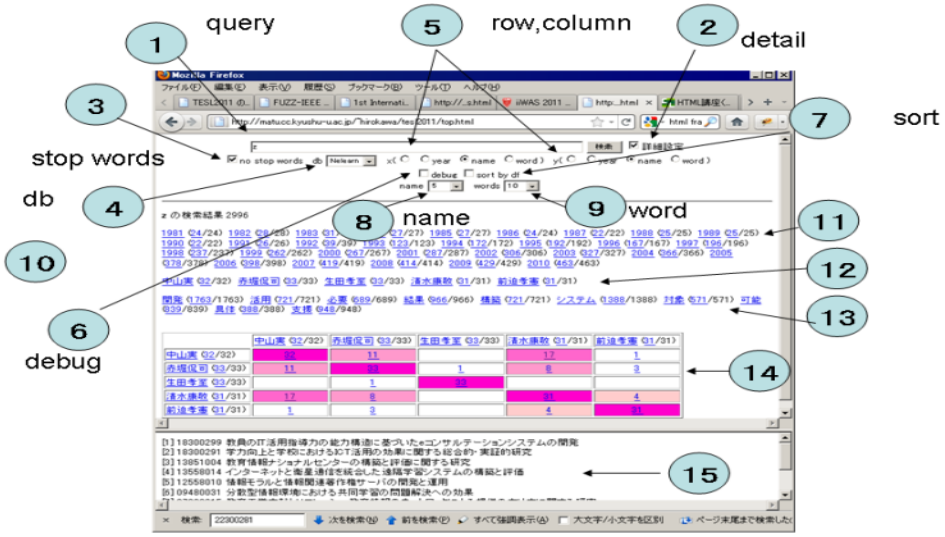

Figure 3. Interface

(b) Output of Features

10. The number of search results
11. The number of documents per year

The fraction shows the number of documents that match the query and the keyword. The denominator displays the number of documents that match the query. The numbers in 12 and 13 are similar. A click on the faction yields a narrowing search using the query and the keyword. A click on the denominator yields a new search with the keyword without the query.

12. *Names*: The name ff researchers are shown for top-N search result. The number is determined in 8 and the order is determined in 7.
13. *Word*: Top-M related words are shown.

(c) Matrix display

14. *Click on cell*: The search result is shown in a matrix, where the x-axis and y-axis are determined in 5 and the number in each cell means the number of documents that match both of x-axis and y-axis. The keywords, years, names of researchers in the first column and in the first line can be used for a new search. As in 11,12 and 13, these meta-data can be used for new search and narrowing search. A click on a cell displays the list of titles of the project in the lower frame.

(d) List of project titles

15. The titles of the projects are shown by clicking each cell. Top 10 titles are shown from the list, even if it contains more.

B. Analysis samples (Hirokawa sensei)

The following scenarios are thought as an analysis procedure.

(1) Choose keywords or names of researcher of your interest.
(2) Choose the attributes from "year", "name" or "word" for x and y-axis.
(3) Search.
(4) Analyze the distribution of projects on the cells and focus characteristic area.

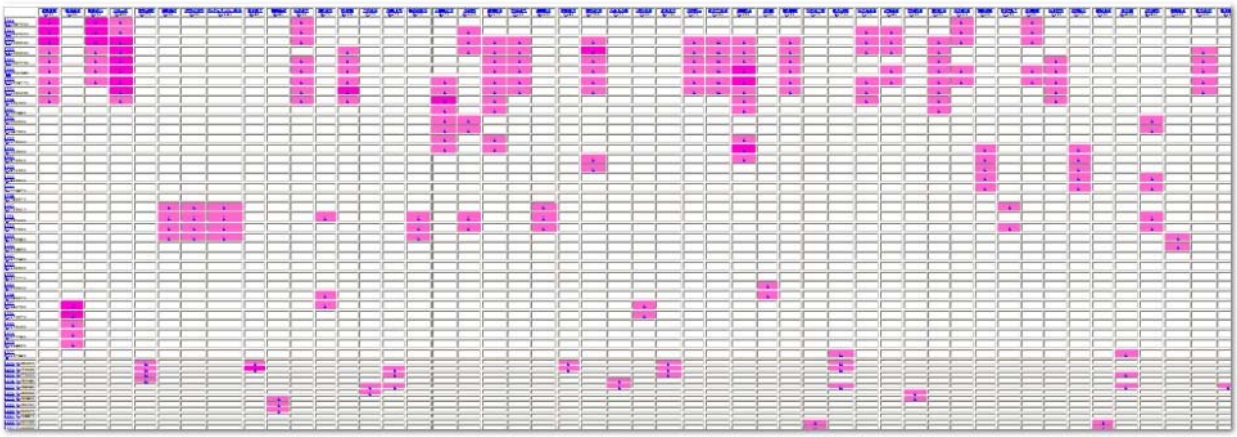

Figure 4. Education and engineering

(5) Change the axis if necessary.
(6) Choose new query from "year", "name" or "word" shown on the x-axis or on the y-axis.
(7) Proceed new search or narrowed search using the query.

It is necessary to input key words at the beginning. However, after this initial query, a more detailed analysis and a new analysis become possible because the user can choose keywords by simply clicking the year, the name or the word displayed on the screen.

Combination of x-axis and y-axis provides several ways of analysis. The user can find who are obtaining research fund for what period by the combination of "name*year". The combination of "name*name" shows research groups. The combination of "year*year" reveals how long projects have been continued and when was the peak of the projects.

In this section, we show a case study with respect to the query "education and engineering". Figure 4 is the map with the axis "year*name", where the names of top 50 researchers are displayed. We observe that projects with a large number of people are executed after fiscal year 2003. The square blocks in the left-lower area represents a project in 1989--1992 and by 5 researchers (Hosota, Mizuta, Quackenbush, Kumatori, Fukada). By clicking the cell, we know the title of the project "Sociolinguistics language technological research on Japanese voice education".

Figure 5 is the search result obtained by narrowed query "education engineering n:TetsuroFurukawa" to analyze top researchers after 2003 in detail. The map was obtained by a click on the number "5" in "Tetsuo Furukawa (5/8)" in the map.

	古川哲朗 (5/8)	山川武人 (4/4)	松石正克 (3/5)	竹俣一也 (4/30)	松本重男 (2/6)	平尾美嶽 (1/1)	星野裕子 (1/4)	向井守 (1/5)	服部隆一 (1/9)	南出重幸 (1/28)
古川哲朗 (5/8)	5	4	3	4	2	1	1	1	1	1
山川武人 (4/4)	4	4	2	3	1					1
松石正克 (3/5)	3	2	3	3	2	1	1	1	1	
竹俣一也 (4/30)	4	3	3	4	2	1	1	1	1	1
松本重男 (2/6)	2	1	2	2	2	1	1	1	1	
平尾美嶽 (1/1)	1		1	1	1	1	1	1	1	
星野裕子 (1/4)	1		1	1	1	1	1	1	1	
向井守 (1/5)	1		1	1	1	1	1	1	1	
服部隆一 (1/9)	1		1	1	1	1	1	1	1	
南出重幸 (1/28)	1	1		1						1

Figure 5. Education engineering n:TetsuroFurukawa

We can see that 7 researchers (Furukawa, Yamakawa, Takemata, Matsuishi, Morii, Minamide, Matsumoto) form a group under the project title "Development of educational program in Asia Pacific region". There are five projects below with Furukawa which are related each other, where are shown as follows.

(1) Surveillance study on coordinated engineering design education network construction in Asia Pacific region
(2) Construction of coordinated engineering design education network in Asia Pacific region
(3) Surveillance study on development of engineering system exchange education program that improves young person's conversation power
(4) Construction of coordinated engineering design education network including intellectual property education in Asia Pacific region
(5) Development of engineering system exchange education program including CDIO process practice that improves young person's conversation power

We have to careful to distinguish "Asai" who has 4 projects are surrounded by the group of "Fururukawa". He is independent to the group and has 4 projects with respect to education and engineering. But, the total number of his projects is 9, which means that he has some other projects. There are 9 titles of his project are shown as follows, where the first 4 are with education and engineering.

(1) "Educational technology" systematization of education and research on the practice
(2) Practical use examination of online CMI for educational practice training that connects directly attached school and educational technology center
(3) Investigation and research on transformation of consideration of curriculum of educational technology education and student
(4) Practical use examination of online CMI for educational practice training that connects directly attached school and educational technology center
(5) Systematization of applied science education in liberal education and research on the teaching material
(6) Research report for synthesis of science education and technical training
(7) Research on integrated systematization of technology education
(8) The curriculum of a consistent life science education and it researches concerning the teaching material small and the junior and senior high school education
(9) Research on establishment of consistent educational system that deepens cooperation of scientific education and technical training

Figure 6 is another map for the search result of "n:HidekiyoAsai" with "year*word" axes. The title before 1982 is "Practical use examination of real-time CMI system that connects directly attached school and educational technology center". The title after 1983 is "Systematization of applied science education in liberal education and research on the teaching material". We can interpret that this researcher expanded his research area in much general situation. The change of his research style is much more

clear if we see Figure 7 with "year*name" axes. He had been conducted single researches before 1986, while he worked as a member of large group.

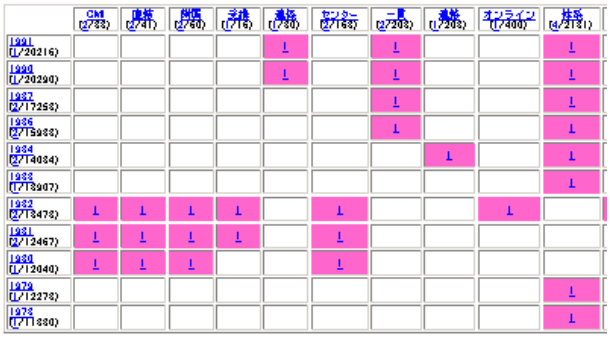

Figure 6. "year*word" axes

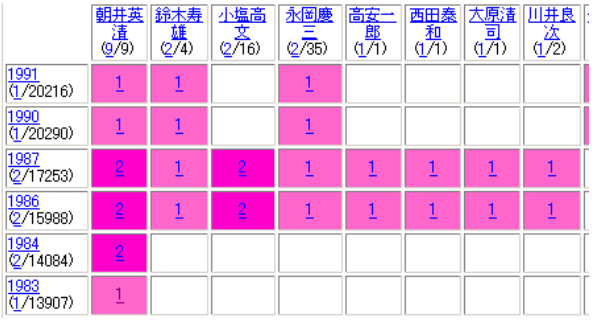

Figure 7. year*name

I. Conclusion and Future Works

In this paper, we advocate learning by searching , which is a method for promoting "active learning" and "discovery learning". online searching is becoming a part of our learning processes, and it is a necessary skill for students. With the development of the Internet and search engine technology, learning by searching will be a very important learning style.

In order to help students make decisions about topics of interest, when they begin to do research. We propose to create a "Learn by Searching" search engine to provide suitable knowledge and help students to master research trends. The KAKEN report is selected as a data recourse, as it suitable to keep up to date with the latest progress.

In the future, we are planning to improve our system to help trend analysis more easily. We plan to conduct an experiment to evaluate the effectiveness of the system.

References

[1] K. K. Bun, Topic Trend Detection and Mining in World Wide Web. Japanese Society for Artificial Intelligence. 2005

[2] S. L. Decker, B. Aleman-Meza, D. Cameron and I. B. Arpinar, Detection of Bursty and Emerging Trends towards Identification of Researchers at the Early Stage of Trends. University of Georgia, Computer Science Department.2007

[3] F. Bouskila and W. M. Pottenger. The Role of Semantic Locality in Hierarchical Distributed Dynamic Indexing.In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI 2000), Las Vegas, Nevada, 2000

[4] G. D. Blank, W. M. Pottenger, G. D. Kessler, M. Herr, H. Jaffe, S. Roy, D. Gevry and Q. Wang, CIMEL: Constructive, collaborative Inquiry-based Multimedia E-Learning.The 6th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE).2001

[5] S. Roy, D. Gevry and W.M. Pottenger, Methodologies for Trend Detection in Textual Data Mining, Proceedings of the Textmine '02 Workshop, Second SIAM International Conference on Data Mining, 2002

[6] H. H. Tsai, Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. Scientometrics. Published online, 2011

[7] H. H. Tsai and Y. P. Chi, Trend analysis of supply chain management by bibliometric methodology. International Journal of Digital Content Technology and its Applications, 5(1), 285–295, 2011

[8] H. H. Tsai and J. K. Chiang, E-commerce research trend forecasting: A study of bibliometric methodology, International Journal of Digital Content Technology and its Applications, pp.101–111, 2011

[9] H. H. Tsai and J. M. Yang, Analysis of knowledge management trend by bibliometric approach. In Proceeding(s) of the WASET on knowledge management ,Vol. 62, pp. 174–178, 2010

[10] Bruner, J.S. (1967). On knowing: Essays for the left hand. Cambridge, Mass: Harvard University Press.

[11] Tseng, Judy C. R., Hwang, G. J., Tsai, P. S., & Tsai, C. C. (2009). Meta-analyzer: A web-based learning environment for analyzing student information searching behaviors. International Journal of Innovative Computing, Information and Control, 5(3), 567-579.

[12] G. Siemens, Connectivision: A learning theory for the digital age. Elearnspace.Retrieved May 18, 2011 from http://www.elearnspace.org/Articles/connectivism.htm, 2004

[13] H. Liu, Learning by Searching. In K. McFerrin et al. (Eds.), Proceedings of Society for Information Technology & Teacher Education International Conference 2008, Chesapeake, VA: AACE, pp. 3843-3844, 2008

[14] P. State, Search engines are source of learning. ScienceDaily. Retrieved May 2, 2011, from http://www.sciencedaily.com /releases/2009/11/091119111417.htm, 2009