A ”milky way research trend” system for survey of scientific literature

Yin, Chengjiu
Research Institute for Information Technology, Kyushu University

Tabata, Yoshiyuki
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 7697 LNCS, pp.90-99, 2014-01-01. Springer Verlag
バージョン：
権利関係：

A "Milky Way Research Trend" system for Survey of Scientific Literature

Chengjiu Yin[1], Yoshiyuki Tabata[1], Sachio Hirokawa[1]
[1] Research Institute for Information Technology, Kyushu University,
Hakozaki 6-10-1, Higashi-ku, Fukuoka, 812-8581 Japan.
{yin, tabata, hirokawa}@cc.kyushu-u.ac.jp

Abstract. Research trend survey is an essential preliminary step for any academic researches, but many beginning researchers have difficulty because they are still foreign to appropriate keywords in his/her research field. We constructed a support system for research trend surveys not only to accelerate the preliminary step but also to let students have a better grips of trend progresses and keyword transitions. Our system assumes a fair amount of data accumulation, for which we employed KAKEN database excerpts, but does not assume manual keyword registration or any other heuristic preprocesses: with an associative search module, it dynamically searches relevant words that are frequently used in the targeted academic field and gives users effective visualizations to understand trend transitions. Preliminary evaluations suggest that the trend transitions that our system presents are effective for trend surveys.

Keywords: Research trend survey, Searching engine, Data mining, Analysis, Scientific literature, Discovery learning.

1 Introduction

With the development of the technology, the longevity of paper literature has become very short. We are constantly required to update our skills and knowledge in order to keep up with technological advancements and meet the needs of scientific research Therefore, it is essential to conduct surveys to have a wide and deep understanding of related research.

Especially for those students who are just beginning to engage in academic research, doing a academic research survey will help them collect the information needed, and guide their planning phases of their projects.

This paper targets on students who are just beginning to engage in research. In order to help students complete a scientific literature survey, with data-mining technologies, using the data of KAKEN [1] (Grant-in-Aid for Scientific Research of Japan), we propose create a search engine to help students to do a scientific research survey.

In 1775, Samuel Johnson said: Knowledge is of two kinds, we know a subject ourselves, or we know where we can find information upon it [2]. This search engine suggests students where to look for solutions to practical problems. At the same time, our system proposes to enable the students to master some of the basic concepts and

methods of scientific literature survey during the process of document retrieval. Students can master research trends through the retrieval results and its analysis.

This research is advocated by pedagogical theories such as 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. Students interact with the world by exploring and manipulating objects, wrestling with questions and controversies, or performing experiments [3].

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. Utilizing this search engine, students can carry out some relevant scientific literature surveys, which 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.

2 Related Works

Previous studies have used content analysis method to identify research trends in e-learning field [4]: Based on the methodology of content analysis, the research topics were first categorized into several tentative categories and sub-categories, and refined manually and continually by using constant-comparative method. By employing scientific papers (abstracts and information) from the five major educational SSCI journals, all those articles are then coded manually to different types of categories referring to its abstract. In addition, highly cited papers are further selected to analyze their research participants, research setting, research design and methods.

Moreover, some researchers used bibliometric methodology to analyze the trends and forecasts in different domains, such as e-commerce, supply chain management and knowledge management [5,6,7]. Using a bibliometric approach, Tsai and Yang analyzed 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[8]. Especially, it used categories such as publication year, citation, country/territory, document types and the like to explore the differences in the two fields.

As mentioned above, these researches require a lot of time to carry out a relevant scientific literature survey. According to statistics, it often costs one-third of the entire time to consult the scientific literature survey in the research process. Doing a research survey is essential, especially for the students who are just beginning to engage in research.

This search engine provides students with a literature survey tool, which not only shows the retrieval results, but also the analysis. Our system does not assume manual keyword registration or any other heuristic preprocesses: with an associative search module, it dynamically searches relevant words that are frequently used in the targeted academic field and gives users effective visualizations to understand trend transitions. This search engine provides a new method to visualize the research trends as "bundles of keywords". We refer to the bundles as "trend milky way".

3 Categories of Published Scientific Literature

A scientific literature survey is a document retrieval method which focuses on literature. It is indispensable for students to master an effective method to carry out a literature survey.

There are many kinds of published literature, such as books, journals, proceedings, sci-tech report. There is some other literature we have not described in this paper such as patent literature. The characteristics of the published literature are as follows (Table 1):

1. Books: Poor time is the problem of with books, it needs a longer period to write and publish. Therefore, books are not suitable to keep up to date with the latest progress. Books are suitable to obtain a general knowledge of a specialized domain, to master the basic content of a specialized problem or method in a short time, to obtain a preliminary understanding of the unfamiliar problem.

2. Journals: Journals focus on a specific discipline or field of study. Journals have characteristics like: 1) strong content innovation, 2) speed report, 3) large amount of information, and 4) it can timely reflect domestic/international science and technology. Therefore, journals are basic form of scientific information transmission and exchanging the academic. Journals are suitable to keep up to date with the latest progress or provide a deep understanding of a specialized field.

3. Proceedings: Generally, conference proceedings have strong academic literature, and it represents the latest achievement in a specialized field. Most of the proceedings are only presented by the results and it is not an inconvenience for knowing specific information. They are suitable to keep up to date with the latest progress.

4. Sci-Tech Report: Sci-Tech Report refers to the government or research departments announced on the official report of research results or actual record of progress during the study phase. The Sci-Tech Report is usually about one year earlier than the journal. It has reported that the original information and results, and had detailed and reliable data. It is suitable to keep up to date with the latest progress. It reflects the National and International trends and technology level.

In this paper, the Sci-Tech Report (The report of KAKEN) is selected as a data recourse, as it suitable to keep up to date with the latest progress and it has detailed and reliable data.

Table 1. Comparison of published literature.

	Latest progress	Detailed Data	Publish speed
Books	×	×	Slow
Journals	Δ	○	Fast
Proceedings	○	×	Very fast
Sci-Tech Report	○	○	Very fast

4 Data-processing

There are three necessary steps for developing a search engine. They are: accumulate data, search algorithm, and provide information. We develop our search engine following these three steps.

Search Engine = Accumulated Data + Search Algorithm + Information Provide

The present paper describes a search engine for project documents by Japanese university researchers. As of May 23 2011, there 74,929 projects are registered in the "KAKEN" database. University researchers can apply their project to obtain research fund from the ministry of education, science and culture of Japanese government. If a proposal is accepted, the researcher can obtain a fund for 2–5 years, depending on their proposal. Those projects are kept in "Kaken" database and are publicly available on the Web. The titles of the projects are listed at the beginning of the project. Progress reports and final reports are shown as short outlines. Each project document contains the following components.

- (a)Identification number of the project
- (b)The name of the project
- (c)The name, Id and affiliation of the principal researcher
- (d) The name, Id and affiliation of project members
- (e)Subject Category of the project
- (f)Keywords
- (g)Duration of the project
- (h)Budget of the project
- (i)Abstract of the project

As a basis of research trend analysis, we constructed a search engine for those project documents, where keywords are extracted from (b), (g) and (i). If a project lasts for several years, for example 1995--1998, the four keywords "y:1995", "y:1996", "y:1997" and "y:1998" are registered as yearly indices.

We employed GETA to realize a search engine dedicated to our system. Information of keyword occurrence as kept as Term-Document Matrix as shown Fig.1. GETA is an associative retrieval engine, specializing in search tasks. It was developed by the Research and Development Center for Informatics of Association, National Institute of Informatics, Japan [9].

The data processing is formed from the following 3 segments:

- (i) Data Collection
- (ii) Construction of a Search Engine
- (iii) Analysis using the Search Engine

"KAKEN" database provides a function to retrieve the project documents by specifying the query. However, it does not provide any tool for high level analysis of the documents obtained. The system proposed in the present paper, initially collets the project documents from the "KAKEN" database (process i). At the next step, we construct a special search engine for the focused documents that are obtained at the process (i). This search engine provides several functionalities for detailed analysis of research trends that can be observed in the target documents.

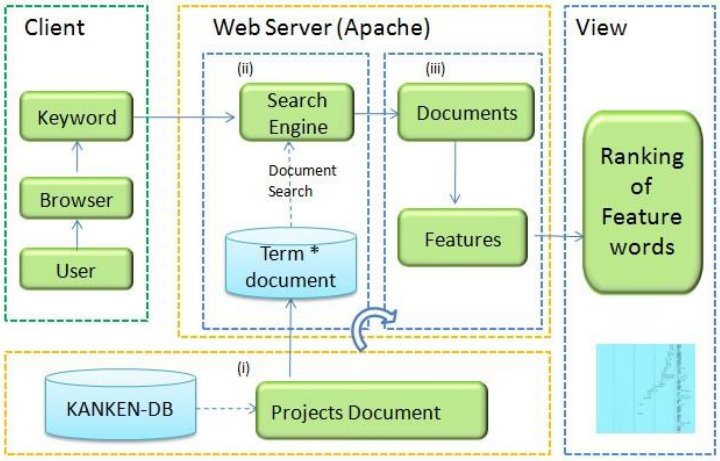

Fig. 1. Search and Feature Extraction.

5 The implementation of the "trend milky way" system.

We used Apache as the server and ran it on Linux, and used Perl to develop the "trend milky way" search engine. As shown in Fig. 2, it is the interface of the system. The learner enters the keywords about his research field and searches for it on the system. A list of the search results will be displayed on the page. The "Research Trend Milky Way" system allows you to search either by “Research Field” or by “Time Range” or by “Sort Key”.

5.1 System Description

1) Advanced Setting. Advanced Setting allows you to change the following features categorized as :

- Search condition: “Research Field”, “Time Range”, “Sort Key”.
- Display Option: “Increase/Decrease Graph”, “OR Search”, “Time Series”, “Left Top->Right Bottom”, “Top ? feature words for each year”, “Total number of results”.

2) Research Field. You can select a research area from the pull-down menu next to the “Research Field”, such as "e-learning", "Text mining". Next to this pull-down menu, there are two time range options. With these two time range options, you can customize the set of time ranges that you view and select from the drop down menu when you search.

3) Sort Key. You can select a sort keyword from the pull-down menu next to the “Sort Key”. There are two sort keywords. One is “weight”, another one is “Frequency” .

4) OR Search. There is a "OR Search" option. OR Search collates the results to retrieve all the unique records containing one term, the other term, or both of them.

The more terms or concepts we combine in a search with OR Search, the more results we will retrieve.

5) Time Series. There is a "Time Series" option, when you choose this option. The following will display a time series analysis graphics.

6) Search. Type the research area of your interest in the textbox next to “Research Field” and click Search Button. The “Research Trend Milky Way” System will display a feature keyword list of related research areas. They are the top slice by keywords frequency or weight. The "Research Trend Milky Way" system allows you to search either by “Research Filed” or by “Time Range” or by “Sort Key”.

7) The number of top ranked feature words for each year. This option means how many top ranked feature words are shown for each year. You can select a number form the pull-down menu next to the “The number of top ranked feature words for each year”.

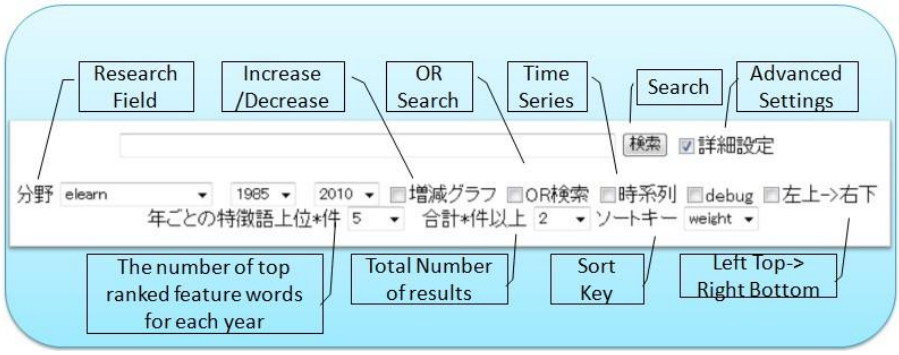

Fig. 2. The interface of “Research Trend Milky Way”

5.2 System Functionalities

Using this system, students can perform trend analysis, automatic extract the outline from literature, and analyze project documents as time-series. This system has the functionalities in the following:

1) Research Trend Milky Way. The Milky Way is drawn as several marks in a plane, where the x-axis designates the year and the y-axis designates a keyword. Imagine that a keyword wj appears as a top-K ranked word in the documents of the year yi. Then a mark "*" is shown at the position (i,j) of the map. Most of the projects last 2 or 3 years, so that the marks appear from left-lower corner to right-upper corner. These occurrences of marks look like a part of the Milky Way. If several projects study the similar subjects, the bundle of keywords that occur in common forms a wider and longer milky way. Two examples are shown in the next proceeding section.

2) Automatic extraction of outline from literature. In order to help students to grasp the outline, problem, method and solution of the literature efficiently, this system provides a method of extracting sentences describing problems automatically from literature abstracts using clue words.

5.3 System Features

There are 3 features of this literature survey 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.

6 Analysis by "Research Trend Milky Way".

6.1 Trend of "Educational Engineering"

The first example concerns 2,886 project documents which contain the keyword "Educational Engineering" from 1998 to 2010. Fig. 3 displays the trend analysis with all of these documents. The system provides several control parameters to choose characteristic features. In Fig. 3, we chose only top 3 keywords for each year and excluded infrequent keywords that appear less than 20 project documents. This graph are basically consistent with Watanabe's investigations [10].

We can see several changes of research trends as follows. The keywords that appeared before 2000, such as CAI(Computer Aided Instruction), Personal Computer, MultiMedia, WWW and Internet, do not occur after 2000. New keywords, such as Distance, Distribution, BBS, Web and Learning appear around 2005. These keywords remind the Web-based distance learning. We can see ICT, SNS, Advancement, Accuracy and Verification as the recent keywords.

We can imagine that educational environment was started using computers in early days, and achieved a progress to distance learning with information sharing between students using BBS and Web. The current ICT trend of SNS influences the educational system as well. A guess with the keywords Advancement, Accuracy and Verification would be that many researchers are considering seriously the evaluation of their system for supporting education.

These observations are just the authors' hypothesis and are not confirmed yet if they are true or not, at the moment. To justify these observation, we need to read the project overview and follow related scientific articles by those researchers. However, this process of generating hypothesis and confirmation of the hypothesis is very important process of study in the era of internet and search engine. The authors think that most of students in young generation are studying this way, in some degree. They use search engines before they read books. They have some impression or hypothesis, which might be wrong after all, before they actually learn something. We think this is how they learn by searching.

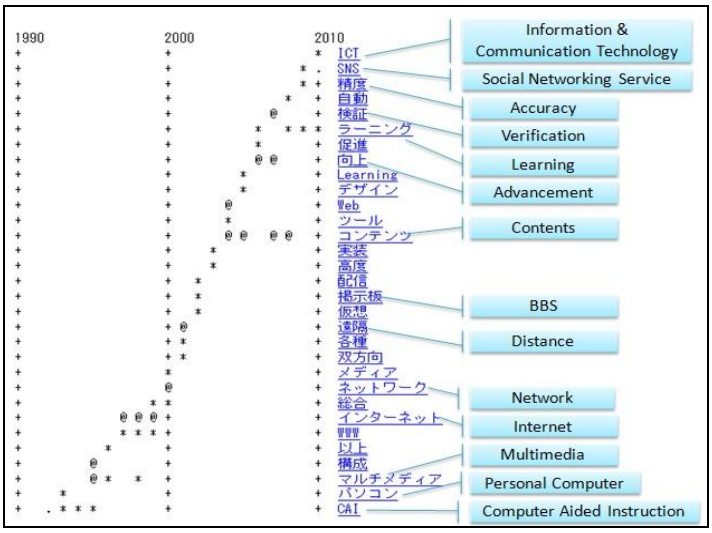

Fig. 3. The interface of "Research Trend Milky Way"[1]

6.2 Trend of "Foreign Language" and "Education"

We constructed another search engine based on 3,034 projects documents that contain "foreign language" and "education". We checked the increase and the decrease of the projects which contain each keyword. It turned out that "English" is among the keywords that increased drastically after 2000. Fig. 4 displays the "trend milky way" of 1,138 documents that contain the keyword "English". This observation is justified by the following information:

> Japanese Ministry of Education issued a statement in 2003 regarding the establishment of an action plan to cultivate "Activities in a foreign language [English])" [11].

We can see the words "elementary school", "junior high-school", "listening" and "guidance" as feature words of the documents around 2005. An implication of this may be that communication skill of English in elementary schools are most active research area in foreign language education. This observation is justified by the following information:

> In 2008, the Japanese Ministry of Education has determined that English be taught in Japanese elementary schools as part of the obligatory educational system. [12]

These two observations are interesting not only in the fact themselves but in the sense that research trends have been greatly influenced by the policies of the government.

---

[1] *The horizontal axis represents years and a list of feature words was shown in the vertical axis. The area was divided by vertical lines which were drawn perpendicular to each other like '＋', these lines represents the years 1980,1990,2000,2010; "@" means that of a frequency greater than 50 times; "*" means that of a frequency greater than 10 times, and less than 50 times;"." means that of a frequency greater than 1 time, and less than 10 times.*

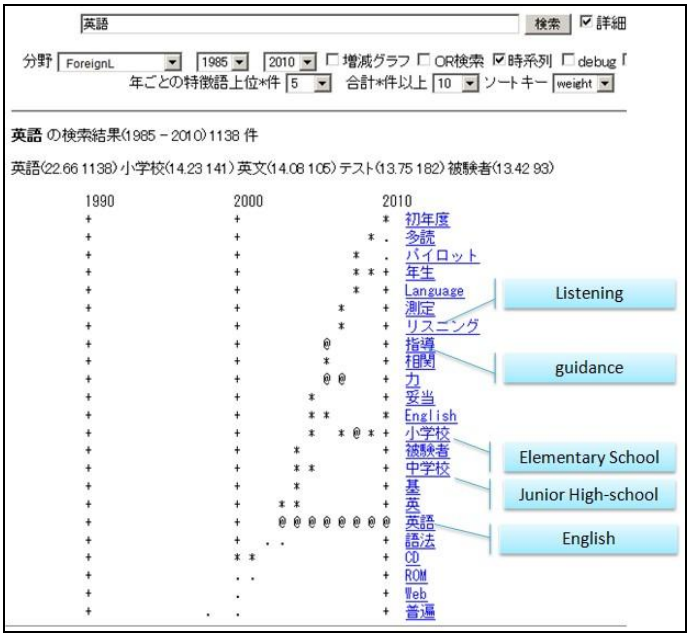

Fig. 4. Trend Milky Way of "Foreign Language Learning"

We constructed another search engine based on 3,034 projects documents that contain "foreign language" and "education". We checked the increase and the decrease of the projects which contain each keyword. It turned out that "English" is among the keywords that increased drastically after 2000. Figure 6 displays the "trend milky way" of 1,138 documents that contain the keyword "English". This observation is justified by the following information:

Japanese Ministry of Education issued a statement in 2003 regarding the establishment of an action plan to cultivate "Activities in a foreign language [English])" .

We can see the words "elementary school", "junior high-school", "listening" and "guidance" as feature words of the documents around 2005. An implication of this may be that communication skill of English in elementary schools are most active research area in foreign language education. This observation is justified by the following information:

In 2008, the Japanese Ministry of Education has determined that English be taught in Japanese elementary schools as part of the obligatory educational system.

These two observations are interesting not only in the fact themselves but in the sense that research trends have been greatly influenced by the policies of the government.

7 Conclusion and Future Works

For the students who are just beginning to engage in research, it is essential for the students to carry out a academic literature survey. In this paper, we propose a system for a research trend survey of scientific literature. With this system, students can perform trend analysis, automatically extract the outline from literature, and analyze project documents as time-series.

We also give some examples to illustrate how effective the system is. We use this system to analyze the trends in the field of "Educational Engineering" and "Foreign Language". Some interesting observations are found such as "research trends have been greatly influenced by the policies of the government".

This is just a prototype system. In the future, we are planning to improve our system to help trend analysis more easily. We plan to analyze other research areas such as data mining, search engines, and then evaluate the results of the analysis by experts/ professors.

References

1. Grant-in-Aid for Scientific Research of Japan (KAKEN), http://kaken.nii.ac.jp/ja/searchk.cgi
2. Johnson S. (1775), Boswell's Life of Johnson, 18th April 1775.
3. Bruner, J.S. (1967). On knowing: Essays for the left hand. Cambridge, Mass: Harvard University Press.
4. Shih, M., Feng, J., and Tsai, C.C. (2007). Research and trends in the field of e-learning from 2001 to 2005: A content analysis of cognitive studies in selected journals [J]. Computers & Education. 51(2): 955-967.
5. Tsai, H.H. (2011). Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. Scientometrics. Published online.
6. Tsai, H. H., & Chi, Y. P. (2011). Trend analysis of supply chain management by bibliometric methodology. International Journal of Digital Content Technology and its Applications, 5(1), 285–295.
7. Tsai, H. H., & Chiang, J. K. (2011). E-commerce research trend forecasting: A study of bibliometric methodology. International Journal of Digital Content Technology and its Applications, 5(1),101–111.
8. Tsai, H. H., & Yang, J. M. (2010). Analysis of knowledge management trend by bibliometric approach. In Proceeding(s) of the WASET on knowledge management (Vol. 62, pp. 174–178).
9. National Institute of Informatics, Japan, http://getassoc.cs.nii.ac.jp/
10. Watanabe, K. & Kashihara A. (2010). A View of Learning Support Research Issues Based on ICT Genealogy. Special Issue: Development of Learning and Educational Technologies on Japan Journal of Educational Technology, Vol. 34, No. 3, 143-152.
11. Japanese Ministry of Education 1, http://www.mext.go.jp/english/topics/03072801.htm
12. Japanese Ministry of Education 2, http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/syokaisetsu/index.htm