# Bibliometric Information Retrieval System (BIRS): A Web Search Interface Utilizing Bibliometric Research Results

**Ying Ding, Gobinda G. Chowdhury*, Schubert Foo, and Weizhong Qian**
*Information Studies Division, School of Applied Science, Nanyang Technological University, Singapore 639798. E-mail: asggchowdhury@ntu.edu.sg*

**The aim of this article is to test whether the results obtained from a specific bibliographic research can be applied to a real search environment and enhance the level of utility of an information retrieval session for all levels of end users. In this respect, a Web-based Bibliometric Information Retrieval System (BIRS) has been designed and created, with facilities to assist the end users to get better understanding of their search domain, formulate and expand their search queries, and visualize the bibliographic research results. There are three specific features in the system design of the BIRS: the information visualization feature of the BIRS (cocitation maps) to guide the end users to identify the important research groups and capture the detailed information about the intellectual structure of the search domain; the multilevel browsing feature to allow the end users to go to different levels of interesting topics; and the common user interface feature to enable the end users to search all kinds of databases regardless of different searching systems, different working platforms, different database producer and supplier, such as different Web search engines, different library OPACs, or different on-line databases. A preliminary user evaluation study of BIRS revealed that users generally found it easy to form and expand their queries, and that BIRS helped them acquire useful background information about the search domain. They also pointed out aspects of information visualization, multilevel browsing, and common user interface as novel characteristics exhibited by BIRS.**

## Introduction

The Internet and WWW have already established themselves as major factors in the operation of scholarly communities worldwide. Today, the Internet is used in all spheres of life for exchange of information. Information resources on the Internet are increasing tremendously. Gordon and Pathak (1999) suggested that the primary use of the Internet is for information retrieval. Search engines are considered as the most important tool for retrieving information on the Web, and consequently form a critical area of research (Gaines, Chen, & Shaw, 1997; Lawrence & Giles, 1998).

Despite the effectiveness of Internet-based or online information retrieval, problems still exist. Woodward (1996) argued that the Internet is currently in a state of near chaos in terms of access and organization of information. Voorbij (1999) found that 67% of the Internet users agree or strongly agree with the difficulty to perform subject searches on the Internet. Users, especially the novice and irregular users, find it difficult to phrase their information needs due to the lack of knowledge literacy in search domain (Bates, 1986, 1998). Although subject headings, traditional thesauri, and term's dictionaries are available to assist users to formulate their queries, their intrinsic shortcomings (out-of-date, inconsistency, lack of search variety, limitation of word's semantic relations—BT, NT, and so on) prohibit them from eliciting user's needs. How to allow searchers who are not familiar with the specific subject area and terminology of a database to express queries using the "vocabularies" from the database or subject area itself has become one of the most pressing questions in the information retrieval field (Chen, Yim, Fye, & Schatz, 1995). These problems coincide with the goals that bibliometric researchers have long addressed in domain-analytic works based on citation data from ISI (Institute for Scientific Information) database, which can also benefit users to form and expand their queries (Nowell, France, Hix, Heath, & Fox, 1996).

Using bibliometric techniques, one can reduce a literature to subsets on the basis of similarity clustering. The criteria generally involve counting the number of times certain markers occur or cooccur, giving rise to information on such author cocitation, journal cocitation, keyword cocitation, and so on. Bibliometric results, especially bibliometric maps, empirically depict states of affairs in various

* To whom all correspondence should be addressed.

Current contact address: ying@cs.vu.nl.

Received 12 November 1999; Revised 23 March 2000; accepted 23 March 2000

fields, such as, the research activity of the prominent researchers, intellectual cluster of leading journals, similarity indication of important themes and concepts, and so on. These maps may be used to assist users to better understand the domain area of interests and address their information needs clearly.

This article reports the development of a Web-based Bibliometric Information Retrieval System (BIRS). BIRS is designed to help end-users formulate and expand queries for searching information on a number of media ranging from OPAC to on-line database and World Wide Web (WWW). The system interfaces with results of an on-going bibliometric research effort in studying the science mapping in the information retrieval field over the 11 years from 1987 to 1997. A number of publications related to this study has been reported elsewhere (Ding, Chowdhury, & Foo, 1999a, 1999b, 2000a).

## Background

### *IR Problem*

With the rapid development of the Internet and World Wide Web, information available throughout the world has increased dramatically. Many information retrieval (IR) researchers are facing some serious problems that they never had decades ago. Three major problems are particularly evident in the context of the electronic age: information overload, query expansion, uncertainty principle.

### *Information Overload*

IR systems search and retrieve data from a collection of documents in response to user queries. The ever-increasing volume of available data is fast rendering traditional IR systems to become less effective. As data continue to grow, it becomes increasingly difficult to develop IR systems that support search and retrieval with little and affordable search times. Although this huge amount of information has posed a great challenge to traditional IR systems, it augments the volume of bibliometric source databases so that more accurate bibliometric mapping results can be achieved (Chen et al., 1995; Fox, Frieder, Knepper, & Snowberg, 1999). If these results can be applied to real information retrieval to yield a bibliometric IR system, then it would have the potential to guide users to form better queries or expand their queries, and at the same time, get a better understanding of specific search domains.

### *Query Expansion*

A major problem in IR is to identify all the items that describe the user's subject of interest. Another problem is to differentiate correctly between the relevant and the nonrelevant documents for that query. The major cause of this stems from the user's inability to use proper words/phrases to form their queries. Many IR researchers have attempted to increase or create more search variety for the end users (Bates, 1986, 1998; Byrne & McCracken, 1999; Chen, Martinez, Kirchhoff, Ng, & Schatz, 1998; Gomez, Lochbaum, & Landauer, 1990; Peat & Willett, 1991). This search variety does not mean the variety in expression, such as word forms (e.g., singular/plural, verb conjugations), syntactical variations (e.g., different word orders), synonymy and so on, but a variety of meaning with similarity. This similarity variety may not be limited only in words, can be extended to any item, for instance, authors with similarity based on author cocitation frequency, journals with similarity on journal cocitation frequency, or articles with similarity based on cocited frequency by other articles.

A user's original query statement typically consists of just a few terms germane to the topic, and it is often necessary to add variety to achieve an effective search. This query expansion process has traditionally been carried out by means of thesauri and controlled vocabularies, which only can provide the variety on the semantic relations of the words. However, maintaining the currency of such tools is extremely expensive and time-consuming. Other broad similarity variety, such as authors or journals, is usually not directly available for the users. Bibliometric techniques, based on the cooccurrence, theory may be one of the best ways to probe the similarity variety of various items.

For similarity variety on words, co-word analysis or term cooccurrence techniques are considered as one of the tools to achieve this goal (Croft, 1995). It is generally assumed that terms used in queries are good at discriminating relevant from nonrelevant document, so that closely associated terms (i.e., terms that cooccur frequently with the query terms) are also likely to be good discriminators. These additional terms may, hence, allow the retrieval of relevant documents that would not have been retrieved using the original query (Salton & McGill, 1983). Numerous co-word and term cooccurrence studies have been conducted to increase search variety for users (Chen & Dhar, 1991; Chen et al., 1995, 1998; Chen & Lynch, 1992; Chen & Ng, 1995; Chen, Ng, Martinez, & Schatz, 1997; Cochrane & Chen, 1996; Peat & Willett, 1991; Schatz, Johnson, Harter, & Cheng, 1996; Seow, 1999; Trivision, 1987). For similarity variety on other items, such as authors or journals, author cocitation analysis, and journal cocitation analysis are proven to be efficient techniques that not only correctly detect rich similarity variety of authors and journals, but also map the dynamic changes over time (McCain, 1984, 1986, 1991, 1998; Tijssen & van Raan, 1990; White & Griffith, 1981; White & McCain, 1997).

### *Uncertainty Principle*

One of the major causes of failure in IR systems is vocabulary mismatch. This means that the information need is often described using different words than are found in relevant documents. At the same time, searchers tend to use different search terms for the same information sought. The average likelihood of any two people using the same word

in their descriptions of the same object is found to range from about 0.07 to 0.18 (Furnas et al., 1982). Both the indexing behavior and information-seeking behavior are at least in part, indeterminate and probabilistic. The indexing inconsistency enhances the uncertainty problem (Bates, 1986, 1998). An individual searcher, however, is usually unaware of the many terms that might be used for search. So the logical strategy in the design of IR systems is to help the searcher generate the search variety (Bates, 1986).

The central problem is that the IR researchers have a hypothesis that the search terms should be in the traditional thesauri or other vocabulary control tools, rather than the database itself to provide the search terms appropriate to user's needs. Researchers have focused on the use of search terms derived from extrinsic sources, rather than from the language used by the database text authors (e.g., Kostoff, Eberhart, & Toothman, 1997).

Bibliometric IR system can solve these IR problems in two ways. First, cocitation analysis (e.g., co-word analysis, author cocitation analysis, journal cocitation analysis) can generate search variety for the users based on the data in the database. The variety provided by cocitation analysis is reasonable and reliable because it is from the data itself and reflects citers' group opinions. Second, cocitation mapping can give users a general view of the entire subject field. Through exploring or browsing such maps, users will be given the prospect in using these maps to identify new avenues for searching.

### *Information Visualization*

In a traditional paradigm of information retrieval, the interaction between users and IR systems relies heavily on a query-based navigation. Visualization can play a pivotal role in reducing the navigational difficulties of traditional IR systems by allowing users to view the structures and processes with which they are dealing. The notion that science can be mapped was first clearly stated by Derek Price during the 1960s (Price, 1966). With the assumption that visualization gives a new insight into IR domain, a set of visual interfaces for IR systems has been developed (Chalmers, 1994; Wise, 1999; Wise et al., 1995), and some of the mapping techniques were compared and updated (Chen, 1999; Chen, Houston, Sewell, & Schatz, 1998; Duplenko & Burchinsky, 1995; Rorvig, 1999). Recent research on information visualization include ISI (Institute for Scientific Information) (Small, 1997, 1998, 1999), CWTS (Center for Science and Technology Studies) (van Raan, 1997), WEBSOM (Kohonen, 1998), SPIRE (Spatial Paradigm for Information Retrieval and Exploration, later on called NewsMaps) (Hetzler, Harris, Havre, & Whitney, 1998), and so on.

### *Bibliometric IR Systems*

Multidimensional scaling maps derived from cocitation techniques reflect natural organizations of the structure of knowledge in both the sciences and the social sciences. Furthermore, they can be useful tools in aiding researchers. Those unfamiliar with a subject can locate the area of interest on a detailed map and obtain bibliometric information for that subject. Even the expert may be led to a related field by the unsuspected proximity of his or her area to another revealed by the map (Garfield, Kimberley, & Pendlebury, 1988).

The idea of setting up IR systems including bibliometric techniques was first tested by Yermish (1975). Later, Nowell and Hix (1993) developed the Envision system that allowed users to control the display of several bibliographic or bibliometric variables during searches. ISI's SCI-MAP system built a cocitation network of documents or journals to assist search (Small, 1994). Hearst's (1995) TileBars aids retrieval decisions by displaying different bibliometric indicators. Butterfly's (Mackinlay, Rao, & Card, 1995) rich 3D display makes visible a basic temporal bibliometric relationship in learned literatures. Wise et al.'s (1995) SPIRE system mapped corpora of related writings on the basis of term cooccurrence statistics and later on updated to a new system called Newsmaps.[1]

Grivel, Polanco, and Kaplan's (1997) HENOCH combined hypertext and database management technologies with the goals of bibliometric analysis. Quoniam, Balme, Rostaing, Giraud, and Dou (1998) presented their system with graphical representation of the results by using bibliometric law. Bollacker, Lawrence, and Giles' (1998) ResearchIndex (later called CiteSeer) uses Web search engines and heuristics to locate and download papers and to extract semantic features, including citations and word frequency information. Fox, Frieder, Knepper, and Snowberg's (1999) SENTINEL system is a fusion of multiple IR technologies including the term cooccurrence and a neural network-training rule.

There are also some other bibliometric IR systems on the Web, for example, BIRD[2] that retrieves a set of similar documents by following citation paths; Alexa Internet[3] that recommends URLs based on the hyperlink paths; Google[4] that uses hyperlinks to identify the most important documents; WebQuery[5] that allows analyze and visualize results of a Web query based on the connectivity of the various documents it contains; WebWatcher[6] that allows users to browse the Web and recommends relevant pages based on an analysis of the hyperlink structure in the neighborhood of pages that the user has visited; Rankdex[7] that indexes Web pages using the anchor text of the various hyperlinks that

[1] Newspaper (URL: http://www.newsmaps.com).

[2] BIRD (URL: http://ai.iit.nrc.ca/II_public/WebBird/index.html).

[3] Alexa Internet (URL:http://www.alexa.com).

[4] Goggle (URL: http://www.google.com).

[5] WebQuery (URL: http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html).

[6] WebWatcher (URL: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/prject-home.html).

[7] Randex (URL: http://rankdex.gari.com).

point to them; Web of Science[8] that offers cited reference searching; HyPursuit[9] that uses a hierarchical document clustering system for the World Wide Web based on analysis of both the word content documents and the structure of the hyperlinks that link them; and Clever[10] that uses hyperlinks to identify the most important documents retrieved by a key words search; and so on.

## Current Bibliometric Research

Bibliometrics is the most useful method for achieving a macroperspective on scholarly communication processes. Cocitation analysis is not only the most important subset, but also the most methodologically sophisticated and popularly utilized method in bibliometrics. Because Small (1973) introduced the concept and defined it as "the frequency with which two items of earlier literature are cited together by the later literature," cocitation analysis has been a potentially productive method and has been successfully applied to examine the cognitive/intellectual structure of scientific specialties, communication patterns of scientific specialties in terms of cocitation among authors, journals, or keywords (Braam, Moed, & van Rann, 1991; McCain, 1984, 1991; Small & Griffith, 1974).

The BIRS system is based on the results of an on-going bibliometric research, which is focusing on science mapping in Information Retrieval area (here, we called it IR bibliometric research). Relevant information retrieval articles were collected from the *Science Citation Index (SCI)* and *Social Science Citation Index (SSCI)* for the period of 1987–1997. In total, 3,325 source articles were selected from 971 source journals with 78,785 citations during the period of 1987–1997. This bibliometric research has been conducted on: authors by using author cocitation analysis (Ding, 1998a, 1998b; Ding et al., 1999a); journals by using journal cocitation analysis (Ding et al., 1999b, 2000a), and keywords by using co-word analysis (Ding et al., 2000b).

### *Author Cocitation Analysis*

Author cocitation analysis (ACA) is a set of data gathering, analytical, and graphic display techniques based on the cocitation frequency of two authors that can be used to produce empirical maps of prominent authors in various areas of scholarship (McCain, 1990; White & McCain, 1998). As a well-established technique in bibliometrics, it is an ideal method to trace the author's role in the scholarly communication process of a specific area. It typically represents authors by the embodiment of their ideas in their whole body of published documents so that the cognitive/intellectual structure of the science specialty can be investigated.

The results of author cocitation analysis in the IR bibliometric research, as displayed by author cocitation maps, have identified the intellectual structures of IR which are subdivided into one "hard" part working on IR theory and retrieval algorithms, and one "soft" part concentrating on the user-system relation (Ding, 1998a, 1998b; Ding et al., 1999a). These maps also depicted the scholarly migration of the prominent IR researchers according to various time periods. The information contained in these maps has a very high potential for the end-users, especially the novices, to gain useful information about various authors (researchers in the field) to form and expand their queries for a real search. For instance, the results can show them:

(1) who are the most prominent authors (researchers) in IR field, what are their highly cited articles, and their research interests, and so on.
(2) Who are the authors having research interests similar to those of the user's familiar authors.
(3) What are the important research groups, or IR subfields, and their relations.

### *Journal Cocitation Analysis*

Scientific journal is an essential component not only to the progress of the intradiscipline as a science but to the interdiscipline as well (Doreian, 1988). Journal cocitation analysis is used to study the structure of scholarly specialties through the published literature focusing on journal similarity based on patterns of citations (McCain, 1991). Journal cocitation analysis provides an operational indicator to investigate the scholarly communication process of specific disciplines (McCain, 1991).

The results of journal cocitation analysis in the IR bibliometric research, as displayed by journal cocitation maps, have detected the interdisciplinary communication among computer science, psychology, physics (optics), chemistry, and science/nature/neuroscience, and IR field; and the intradisciplinary communication within the IR field (Ding et al., 1999b, 2000a). These results or journal cocitation maps can help the end users get a better understanding of the IR field, and therefore, accurately formulate their information needs while searching. For instance, these results can show the end users:

(1) What are the highly cited journals in the IR field, and their research focuses?
(2) What are the core journals in the IR field?
(3) What journals have research interests similar to those of the user's familiar journals.
(4) What are the inter- or intradisciplinary relations of the IR field.

### *Co-word Analysis*

In a scholarly communication process, the content of information can be communicated as a concept, idea, or

[8] Web of Science (URL: http://www.isinet.com).

[9] HyPursuit (URL: http://paris.les.mit.edu/Projects/CRS/HyPursuit).

[10] Clever (URL: http://www.almaden.ibm.com/cs/k53/clever.html).

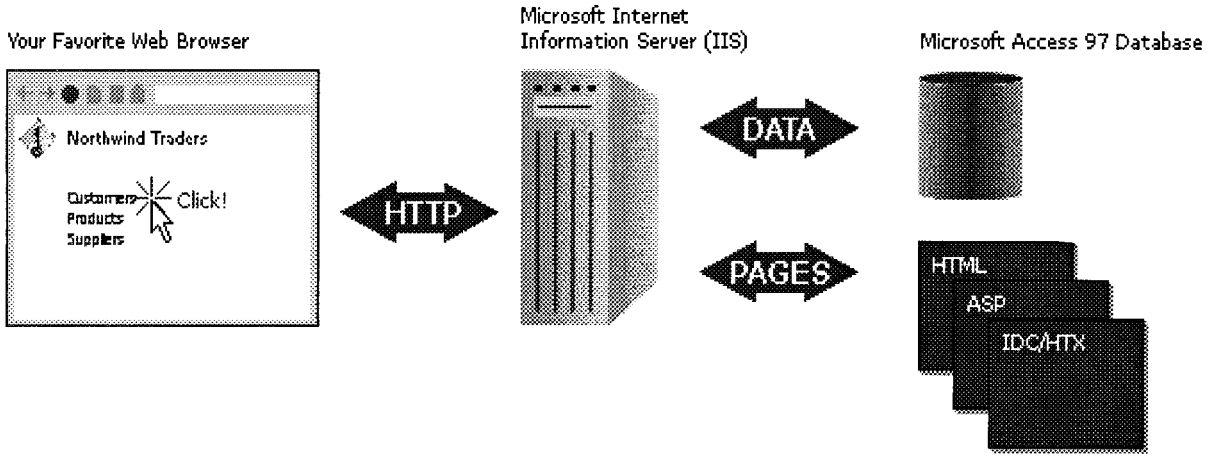


FIG. 1. Microsoft Web database application solution.

knowledge. The content transferred in the scholarly communication process can be monitored by co-word analysis in terms of the keywords or phrases used. Co-word analysis is delineated as a research technique for the objective, systematic, and quantitative description of the manifest content of communication by identifying keywords representing their research content and linking articles based on the term cooccurrence to produce an intellectual map of a specialty (King, 1987).

The results of the co-word analysis in the IR bibliometric researches, as displayed by the co-word maps, have traced the knowledge transferred in the IR field. This field has some established research themes (user study, IR model, IR theory, and so on), but it also changes rapidly to embrace new themes (Internet, Web search engine, natural language processing, and so on). The role of implicit subject keywords or newly emerged subject may be crucial in observing trends between and across updates to the IR area. These co-word maps not only can give the end users a detailed overview of the IR field, but also furnish similarity relations of the words or phrases that are distinct from those semantic relations provided by traditional thesauri so that more search variety can be generated for the end-users. For instance, the co-word results can tell the end-users: (1) what are the important research topics in the IR field and their interrelations; and (2) which keywords have high cocitation similarity.

One of the major objectives of this research was to test whether the results of this on-going bibliometric research in IR can be used to help end users in their search and query formulation and expansion activities. This study aimed to develop the BIRS interface that allows users to recall, display, and use the authors, journals, keywords, and corresponding cocitation maps in course of their search processes.

## BIRS System Design

The BIRS uses a Microsoft MS-Access database to store records of the important authors, journals, and keywords, and bibliometric maps that encompass author cocitation maps, journal cocitation maps, and co-word maps. These maps were linked to the database by using JavaScript, VBScript, and ASP (Active Server Page) to form the BIRS system. The characteristics and system design of BIRS is elaborated below.

### *Hardware/Software Environment*

The BIRS is designed, implemented, and maintained in an environment running Microsoft Windows 98/NT operating system. The WWW server of the BIRS is Microsoft Internet Information Server (IIS) on Windows NT 4.0. In the BIRS interface design, Microsoft Frontpage 98 was used as the HTML file editor, and Microsoft Visual Studio 5.0 as the project editor. PWS (Personal Web Server) was used as the Web Server, which is the desktop version for IIS (Microsoft Internet Information Server). Microsoft Access 97 was used as BIRS database, ODBC server as the connection between web application and database. For programming, ASP (Active Server Page) was used as server-side scripting, which embedded with HTML, VBScript, and JavaScript; VBScript, and JavaScript as Client-side scripting (Powers, 1998).

### *Design Features of the BIRS*

The typical model for the Microsoft Web database solution is shown in Figure 1. A user uses the Web browser to make a request to IIS for the ASP file. The IIS reads the ASP file, which contains the SQL statement(s), connection information, and field value placeholders. Subsequently, the VBScript/JavaScript code opens the Microsoft Access database (using the Microsoft Access Desktop ODBC Driver and the ASP file connection information) and runs the query in the ASP file to access the data. IIS sends the HTML file back to the Web browser for display as a Web page. The BIRS system uses the Microsoft Web Application Model to connect with the Access database and guides the users to

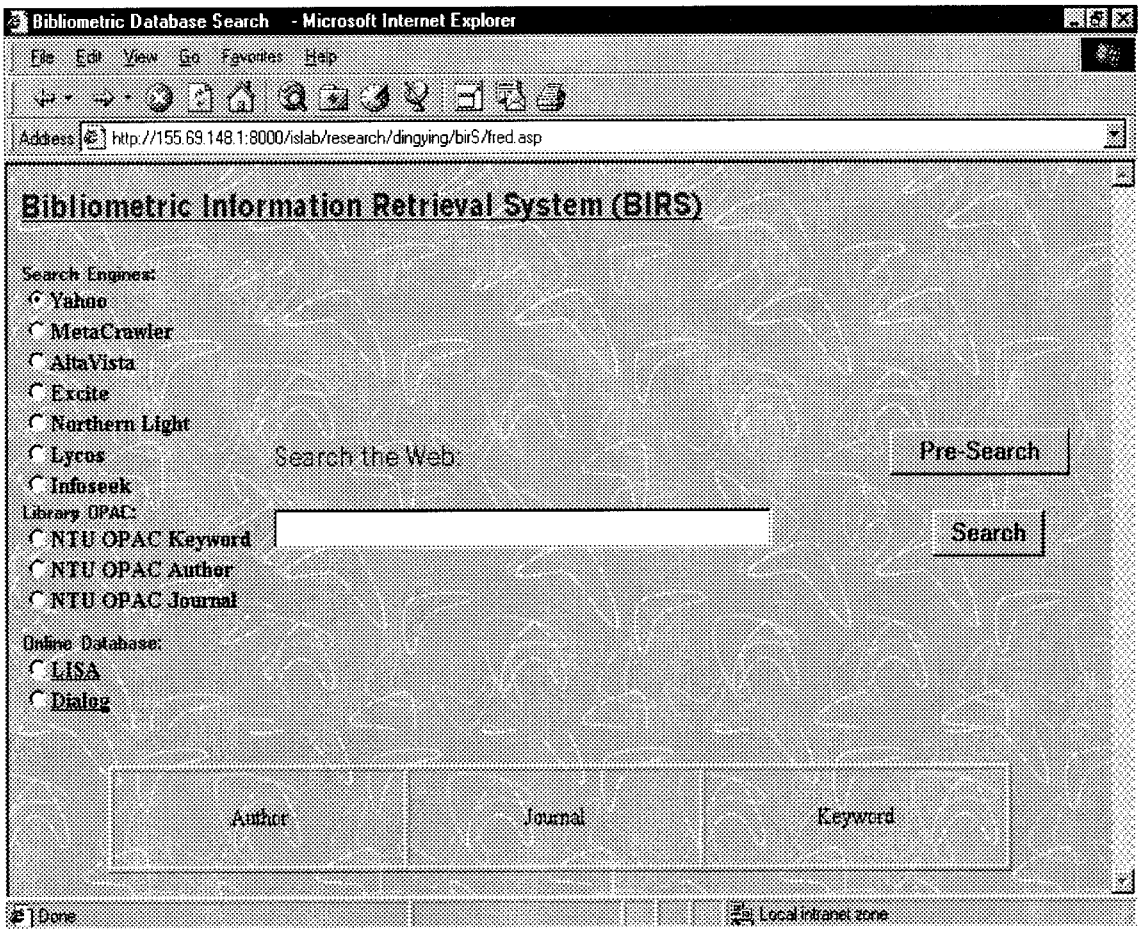


FIG. 2. The BIRS homepage.

browse the bibliometric results in order to help them on the subject search.

The BIRS is designed and implemented as a software system to expand and refine user's query with a rich set of features in an efficient, easy, and globally accessible way. The BIRS is a globally available tool on the WWW with a consistent, visualized user interface, and multiple searching choices.

BIRS has been connected to three kinds of search engines: a range of widely used Web search engines; library OPAC (that is currently the Nanyang Technological University's (NTU) Library OPAC) and on-line database (see Figure 2). Two search options, namely, Presearch and Search, are available to the user. The Presearch option activates the main BIRS interface, as shown in Figure 3, to allow the user interact with the bibliometric results. The Search option links the completed query to the selected search engines.

From the BIRS main presearch page, the user can choose to approach their search by author, journal, or keyword, as well as selecting the time span of the search (i.e., 1987–1991, 1992–1997, or 1987–1997). According to the bibliometric research results, the author and journal search can be further broken down into two subcategories: VAR (author or journal information from various fields that encompas library and information science, computer science, and other fields that are related to IR) and LIS (author or journal information solely from library and information science field).

The screen layout of the main search page groups together all the three categories and their subset search option into one page for ease of use. The BIRS main search page consists three frames: left frame (items to be populated), right-up frame (categories option), and right-down frame (category content). User can choose any category (author, journal, keyword, type, and year) in the right-up frame, and after clicking the search button, the relevant items will be displayed in the right-down frame. If the user is interested in a particular item, the hyperlink will lead the user to the details of the selected item. At the same time, these (i.e., author name, journal name, keywords) are automatically listed in the corresponding list box in the left frame. The user can choose and highlight the relevant items to refine and expand the original query (Fig. 4). Thus, this interface can be used to gradually build up a list of search terms associated with author, journal, and keyword categories. With the selected items, the user returns to the homepage to further refine the query using the available Boolean operators (OR, AND, NOT) (Fig. 5).

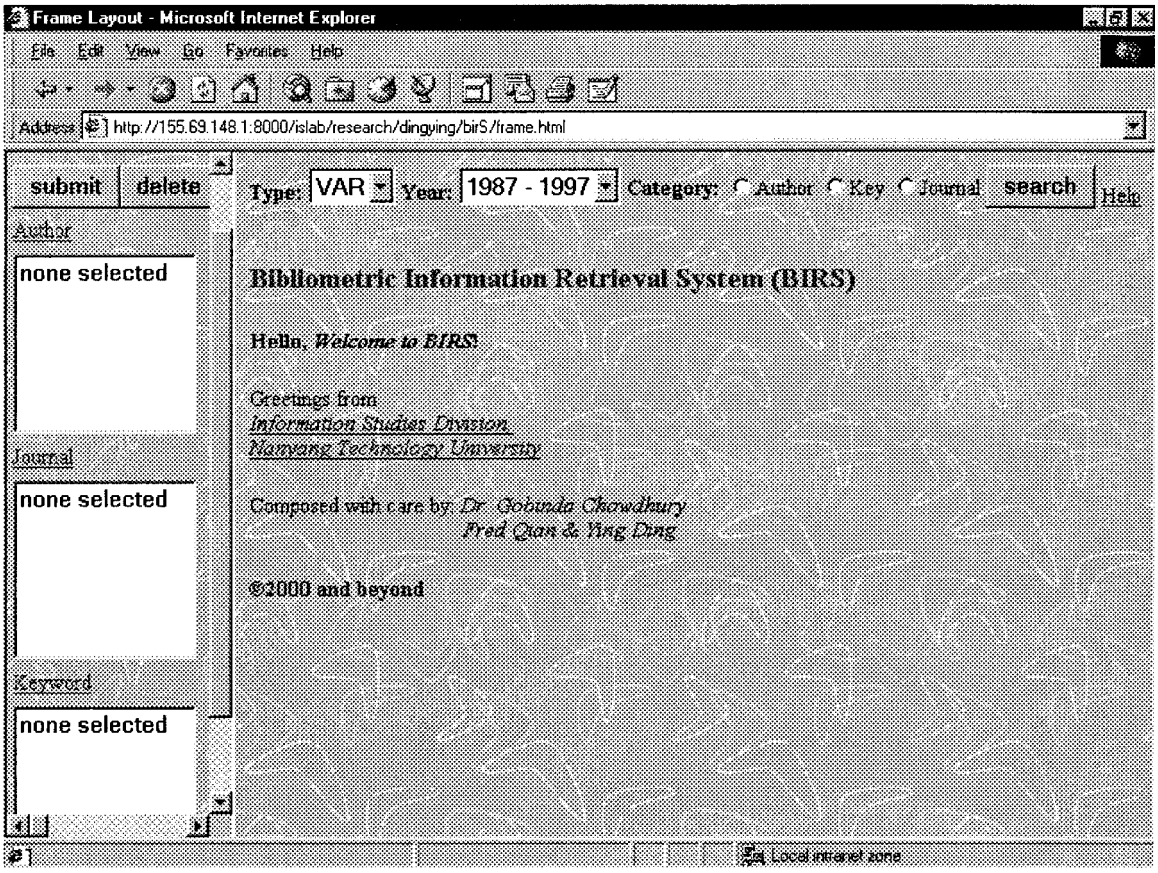


FIG. 3. BIRS main search page.

Finally, the user selects the search engines for conducting the search using the refined query. The process of using BIRS to carry out a search is outlined in Figure 6.

BIRS exhibits three specific design features, namely, information visualization, multilevel browsing, and use of a common user interface.

*Information Visualization.* The maps serve the dual purpose to assist the user in better understanding the search domain and in forming the relevant search expressions. For example, in the author map of Figure 7, users can easily identify the important research groups (clusters) in the IR field. Each author on the map, once clicked, will yield detailed information about the author, including the highly cited articles of the author, published journals, volume, and issue numbers, the number of times the article is been cited by other researchers, and so on. Similarly, the journal map (Fig. 8) can be used to identify the important journals that have reported research findings in the IR field.

In the keyword maps (Figs. 9 and 10), users can obtain general information about the IR field via the overview map. Once they go deeper to the selected cluster, they will be provided with detailed information about the subdomain, such as the intellectual location of specific subject, the relationships of different subjects, relevance of different subjects, and so on.

*Multilevel Browsing System.* Multilevel browsing is incorporated into BIRS to support layering so that users can slice and dice to get different levels of information about interesting topics (Dillon, 1992; Ingwersen, 1984; Saracevic, Mokros, Su, & Spink, 1991). For example, three levels of details are available for the keyword map as shown in Figure 11. The top level (level A) offers an overview of the IR field. Clicking on a cluster results in a more detailed map of the specific cluster (level B). Clicking on an appropriate keyword results in the 20 most relevant associated keywords (level C).

*Common User Interface.* A common user interface enables end users to search all kinds of databases regardless of different searching systems, different working platforms, different database producer and supplier, such as different on-line search engines, different library OPAC systems, or different local or on-line databases. Thus, a common user interface will result in greater synergy between IR systems and end users, or even providers, distributors, and information professionals. The importance of the common user interface has been highly emphasized in many works (Bradley, 1995; Chowdhury, 1999; Chowdhury & Chowdhury, 1999). In BIRS, the common user interface is used to link seven different Web search engines, one university OPAC system, and two on-line database systems together (Fig. 2).

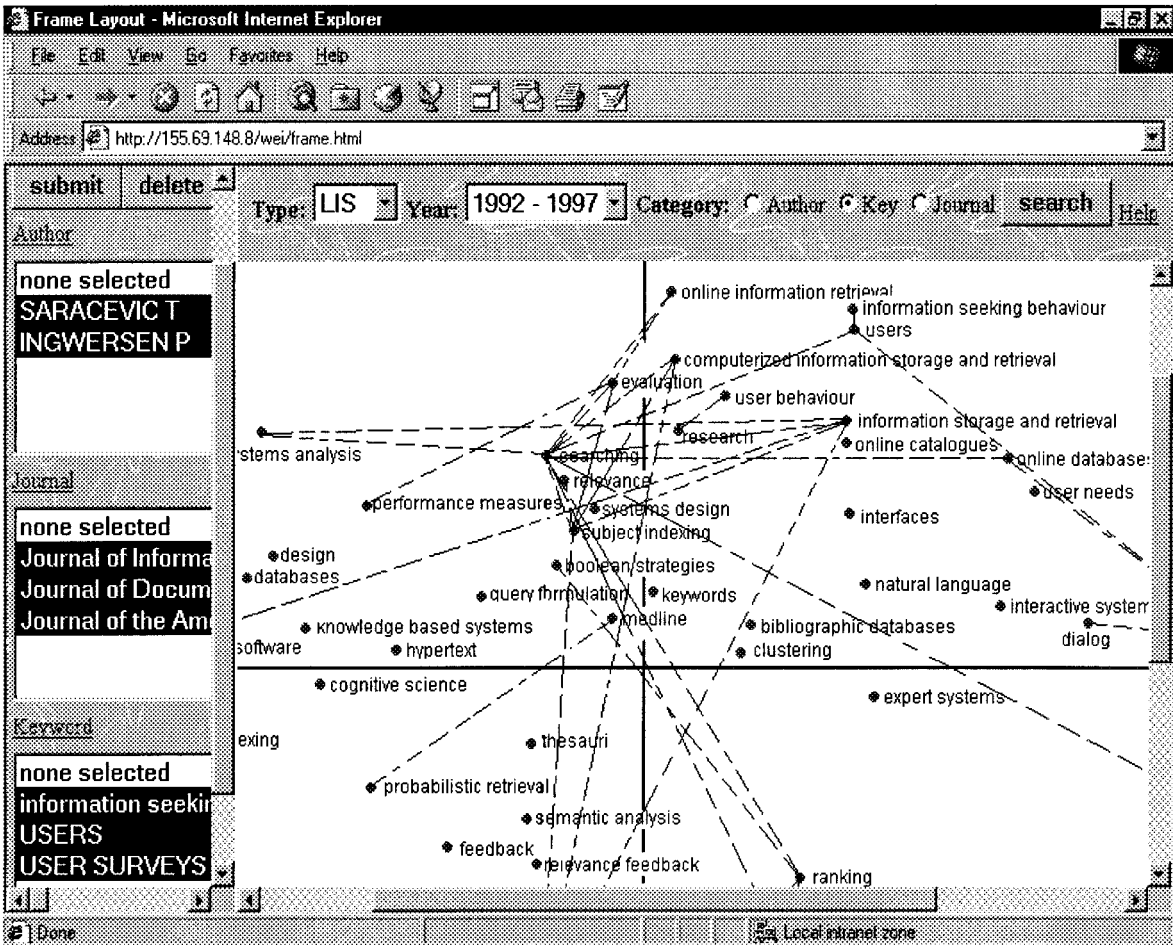


FIG. 4. Example of different contents in three frames.

This list may be expanded and additional information resources can be linked as and when necessary.

## User Evaluation

A preliminary evaluation of BIRS was conducted in summer of 1999 with 35 subjects at the Nanyang Technological University (NTU). The evaluation aimed to address the following research questions:

(1) Does BIRS help users form their queries?
(2) Does BIRS help users expand their queries?
(3) Does BIRS help users better understand their search domain?
(4) Does the BIRS's common user interface provide a convenient way to help users search for information from a variety of sources, including the Web, library OPAC, and on-line databases?
(5) Are the information maps of BIRS helpful in guiding the search process?

### *Experimental Design*

The experiment consisted of two parts: searching without BIRS, and searching with BIRS. It was carried out in the following sequences:

(1) The subjects were asked to choose one search topic from the six topics provided in the questionnaire.
(2) The researchers (i.e., the authors) provided an overview of BIRS, highlighting the basic theory, important functions of the system, and the procedure for using the system.
(3) The subjects were asked to conduct the search without BIRS. They needed to form or refine their own queries based on the their own understanding of the chosen search topic. They selected a Web search engine, library OPAC, or on-line database for the search. The first 20 records of the results were retrieved.
(4) The subjects made their relevance judgments as "Relevant" and "Not Relevant." For each record in the result set, judgments were made to indicate the relevance of that record to the search question. Relevance was not defined, but was left to individual interpretation by the subjects.
(5) In the second part of the experiment, the subjects conducted their search on the same search topic using BIRS. They familiarized themselves with BIRS's functionality. They used BIRS to get a better understanding of the IR domain, reformed and expanded their queries by choosing relevant authors, appropriate journals and keywords with high similarities.
(6) The subjects subsequently went back to use the same Web search engine, library OPAC, or on-line database

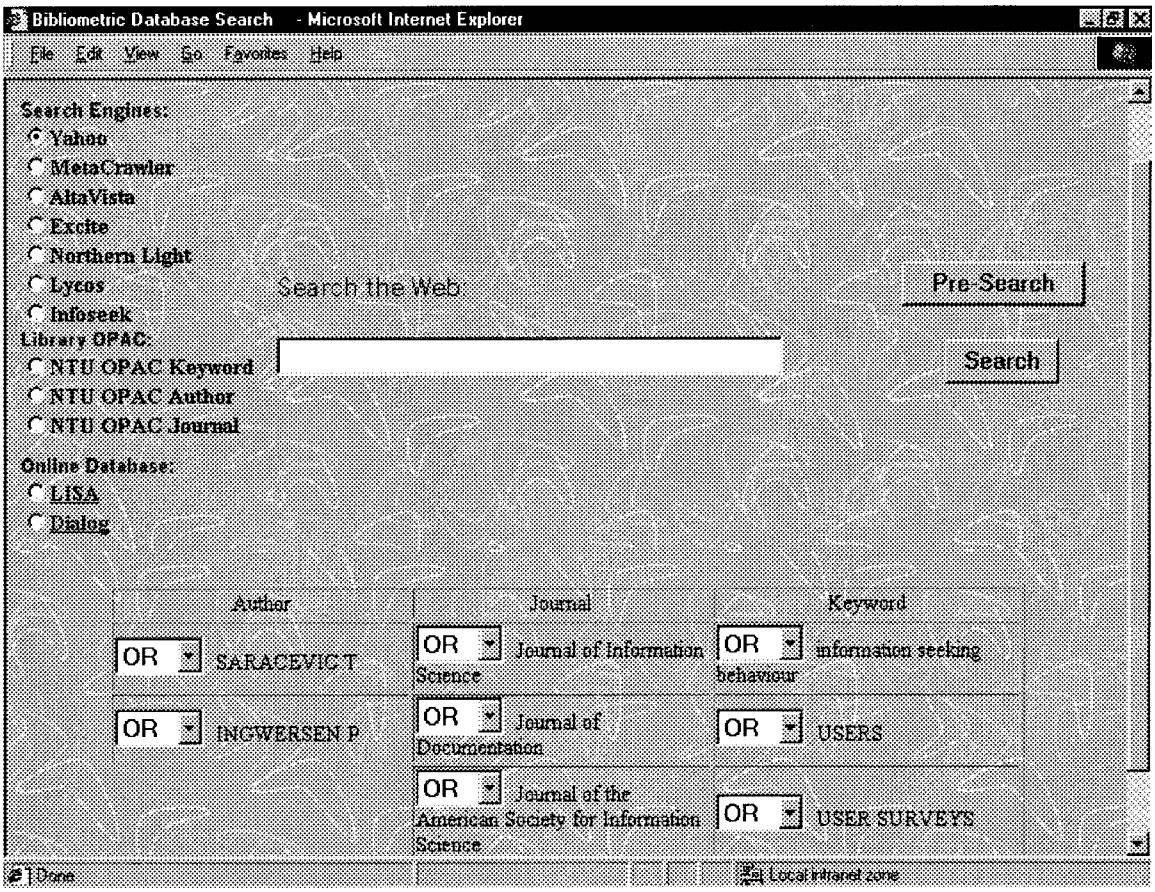


FIG. 5. Expand and refine the query in the BRIS homepage.

to conduct the search again, but based on the reformed or expanded queries. The same relevance judgments were made on the new results based on the expanded queries.

(7) The subjects were asked to compare the two sets of results to judge whether the results were improved based on the expanded queries. They were also asked to comment on BIRS. In particular, they were asked about likes, dislikes, ideas for improvements, such as which set of results they are satisfied with and whether BIRS did help them form and expand their queries, and what are the problems of this system and how to improve them.

### *General Information About the Subjects*

Twenty-nine subjects are postgraduate students undergoing their higher degree in Information Studies. Among them, 24 were second-year students undergoing their Masters degree in Information Studies, and 5 were Ph.D. students in Information Studies. The remaining six subjects are employees of IT-related companies. All used Internet very frequently, and have more than 2 years of Internet searching experience. They are familiar with the popular web search engines like Yahoo, Infoseek, Alta Vista, and so on. They are also familiar with different search environments, such as Web search engines, library OPACs, and on-line databases.

When they began to perform a subject search in the experiment, 11 (31%) subjects did not have a very good understanding of the search domain, 8 (23%) subjects could not find proper words to form their queries, and 19 (54%) subjects could not find proper words to expand their queries. These are common searching problems, which have been described and reviewed in the literature (Bates, 1986, 1998).

The subjects reported that they would like to have an online interactive system that can:

(1) help them better understand their search domain (27 subjects, 77%)
(2). help them form their queries (30 subjects, 86%)
(3) help them expand their queries (30 subjects, 86%)
(4) provide information about famous researchers or important journals in the search domain (31 subjects, 89%)
(5) provide group of keywords that are related to the chosen keyword (30 subjects, 86%)
(6) help the user send a query to any system (Web search engine, OPAC, and on-line database) from the same interface (27 subjects, 77%)

### *User Feedback*

Major points that emerged from the subjects' feedback are as follows:

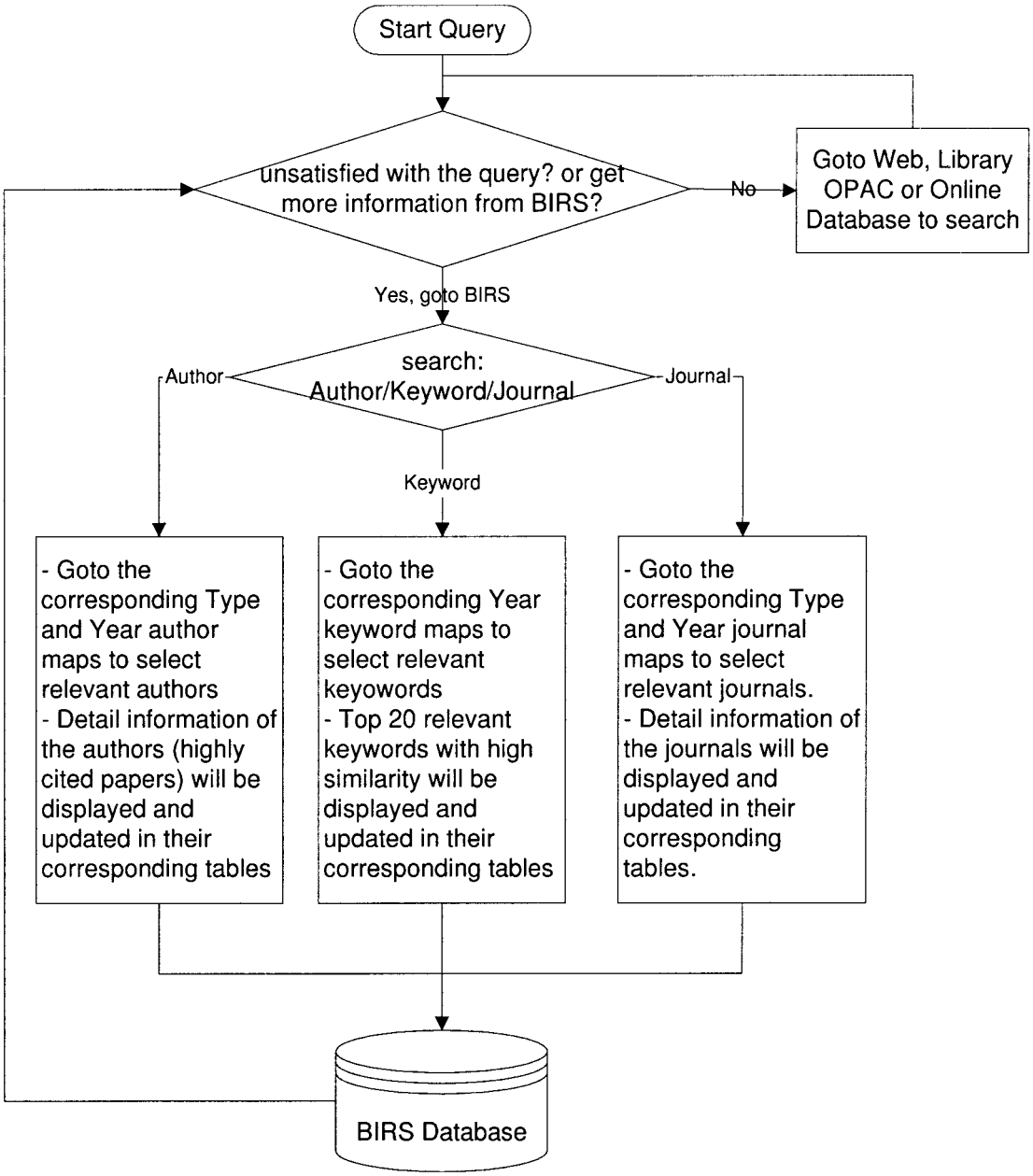


FIG. 6. The BIRS flow chart.

1. *Understanding of the IR area*: 28 (80%) subjects got a good or very good understanding of the IR area with the help of BIRS; 7(20%) subjects did not improve their understanding of the IR area by using BIRS; 31 (89%) subjects chose the Keyword aspect as the most useful part in getting a better understanding of the IR area.
2. *Form and expand query*: 27 (77%) subjects agreed that the BIRS system can greatly help them form and expand their queries; 5 (14%) subjects were neutral, and 1(4%) subject agreed partially; 32 (91%) subjects chose the Keyword aspect as the most important part to help them form and expand queries; 12 (34%) subjects selected the Author aspect, and 10 (29%) subjects selected the Journal aspect as necessary compliments to help them form and expand their queries. In the Keyword aspect, 25 (71%) subjects could find keywords located near their search terms according to the keyword maps, and 21 (60%) subjects could find the keywords with high similarity with their search terms according to the top 20 relevant keyword lists.
3. *Multilevel browsing system*: 25 (71%) subjects indicated good satisfaction with the multilevel browsing system; 9 (26%) subjects were neutral; and 1(3%) was partially satisfied.
4. *Information visualization*: 28 (80%) subjects gave good or very good comments on the helpfulness of the information visualization feature of BIRS. Subjects liked the maps, and could quickly determine which area of the map had the most interesting topics; and that they could zoom in the special areas for closer inspection. Three

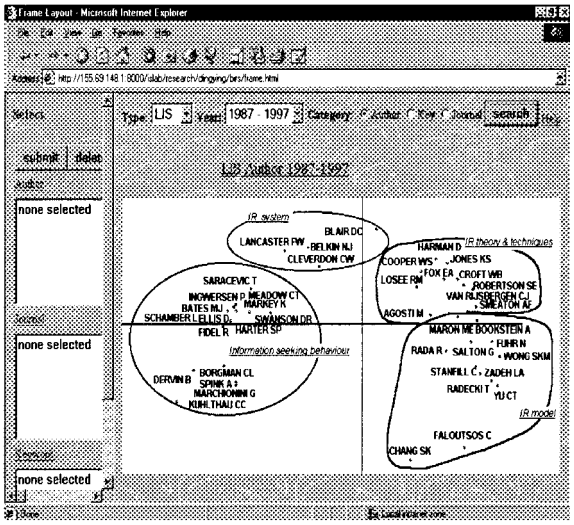


FIG. 7. Author map.

(9%) subjects were neutral on this aspect and 1 (5%) subject thought the information visualization feature was only partially helpful.

5. *Common search interface*: 24 (69%) subjects thought it was easy to follow; 11 (31%) subjects thought there was no need to know different search engines; 12 (34%) subjects thought there was no need to go to different search engines, library OPAC or online databases to search; 12 (34%) subjects thought it can save time; and 11 (31%) thought it was easy to compare the results from different search engines.
6. *Search query improvement*: Using BIRS, 15 (43%) subjects added new author names, 9 (26%) subjects added new journal names and 22 (63%) subjects added new keywords to expand and refine their queries, while these subjects have experienced problems to form their queries before using the BIRS system:
   (a) *Authors:* the author names selected from the author cocitation maps could be more useful in an library OPAC and on-line database (e.g., LISA or Dialog), than Web searching, because users can get the newly added authors' published articles or books directly from the library OPAC and on-line database while they are unlikely to get anything relevant from the Web unless these authors' homepages are publicly available. However, once users can find these authors' homepages, they can capture other valuable information, such as, his/her biography, research interests, affiliated research groups, current projects, personal academic links, and so on. In the current evaluation, 15 subjects added new author names to their refined queries, among them, 8 added one new author name, 5 added two new author names, and 2 added three new author names.
   (b) *Journals:* journal cocitation maps are good at introducing journal's name, function, and publication scope to the end-users, especially novices. If users add new journal names to their refined queries, they can acquire information, such as, the abstracts of the articles published in this journal from online databases (such as LISA or Dialog) or Library OPACs. Sometimes, if luckily, this journal's homepage is

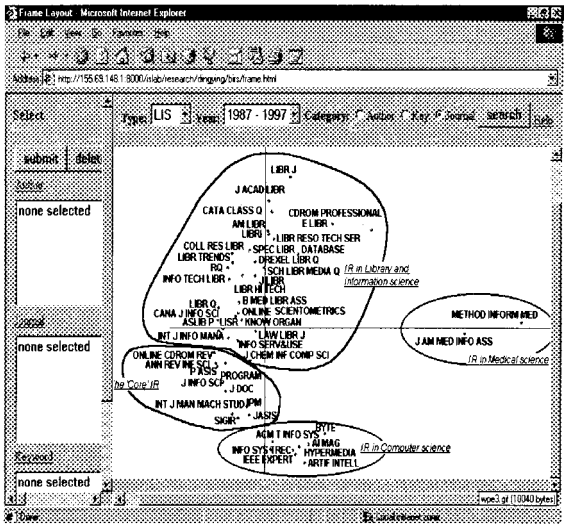


FIG. 8. Journal map.

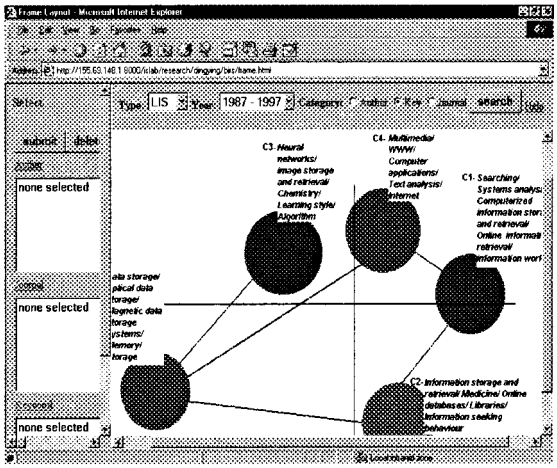


FIG. 9. Overview keyword map.

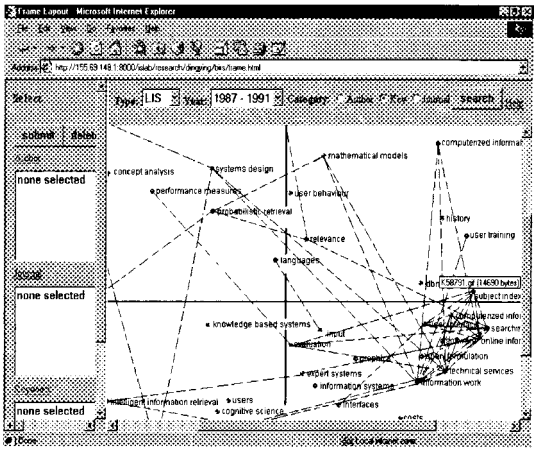


FIG. 10. Detail keyword map.

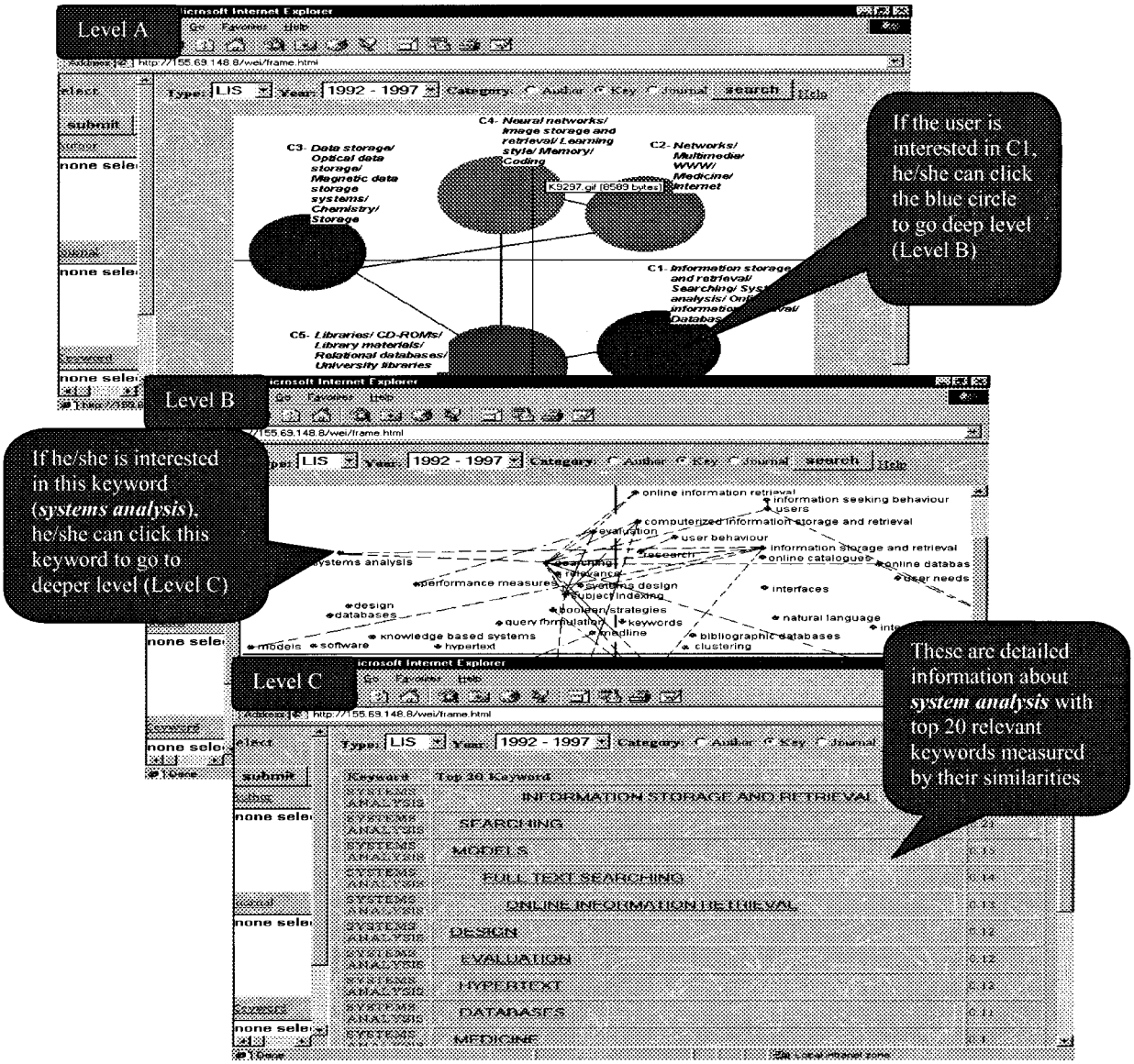


FIG. 11. Multilevel browsing feature of the BIRS.

available, they can gain other additional information, such as journal's editorial policy, call for papers, author's notice for publication, other useful links, and so on. In the current evaluation, nine subjects added new journal names to refine their queries, three added one new journal name, five added three new journal names, and one added four new journal names.

(c) *Keywords:* almost all subjects (91%) highlighted the keyword aspect as the most helpful tool to form and expand their queries. It is understandable because keyword searching is always the most frequently used search method in all types of IR system. It is also the basic unit for indexing and cataloging, and basic search unit in full or free text retrieval systems. In the current evaluation, 22 subjects added new keywords to their revised queries. The numbers of newly added keywords differ from 1 (five subjects) through 7 (one subject). Most of the 22 subjects added one to three new keywords to their new queries (see Fig. 12).

### *Discussion*

The evaluation pointed out problems of BIRS in two major areas: system problems and user interface design issues. On the system problems, some subjects criticized that BIRS was not real-time or on-line but static ("How come these maps are same when I conduct a different search?"). Some subjects could not find any results from some Web search engines based on the expanded queries because they all chose AND as the Boolean operator that made their queries very restrictive. Some subjects commented that "choosing the relational keywords, AND, OR, NOT must be done carefully. This depend on the user's knowledge of choosing the correct combination," and that the "the expanded queries may be meaningless after choos-

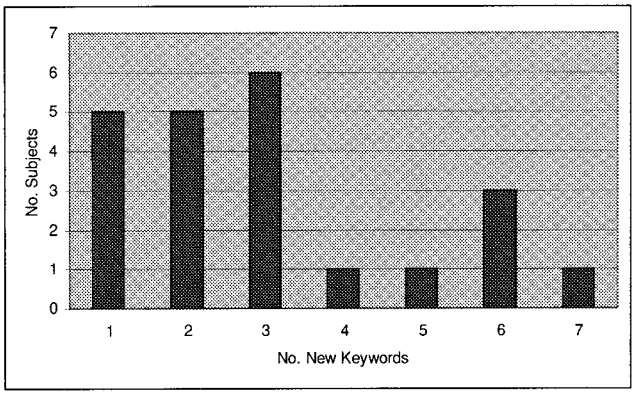


FIG. 12. The distribution of new keywords adding to refine the queries.

ing all the added items for the BIRS with AND, OR, NOT operations." In this instance, we only aimed to test whether the basic ideas of this system can help user form and expand their queries. Choosing the Boolean operator, and its combination, is based on the user's understanding of the search questions and expertise in the search process. It is a common problem that if a query is restrictive, then either the right information is obtained or no information is obtained. The users' original queries are also very important because many researches noted that query expansion led to the greatest improvement in performance when the original query gave reasonable retrieval results, whereas expansion was less effective when the original query had performed badly (Frants, Shapiro, Taksa, & Voiskunskii, 1999; Peat & Willett, 1991).

On the user interface design issues, some subjects complained that it was difficult to understand some maps ("There were too many items in some maps and there was no way to scan them quickly"). Some subjects tended to get lost or confused ("How do you determine what level of the map you are on?," "Am I in the author map or others?," and "I forget where I am and how I got here"). Difficulty to understand the maps for nonbibliometric researchers is a common problem in bibliometric research area (Noyons, Mode, & Luwel, 1999).

In summary, this preliminary evaluation indicated that BIRS was found useful in assisting query formation and expansion, and provided a useful means to acquire background information about the domain area in one integrated system. The information visualization, multilevel browsing, and common user interface are also deemed as novel characteristics of BIRS.

## Conclusion

Users of information retrieval often face the critical problem to form and expand their queries (Bates, 1986; Chowdhury, 1999; Peat & Willett, 1991; Voorbij, 1999). We have attempted to incorporate and integrate the results of an on-going bibliometric research to form BIRS to help users in query formulation and expansion, and to help them acquire new knowledge about the domain. The results of the user evaluation of the BIRS confirm that this system can help user form and expand their queries as well as aid users to better understand the information retrieval domain area. User feedback also clearly indicates that users like the graphical nature of information organization, multilevel browsing system, and common search interface.

As this is a first version of BIRS, many areas need further refinement, enhancing, and development. It is also undeniable that BIRS needs to be extended to cover larger subject domains to make it more useful to a wider community of users. With the availability of results of on-going bibliometric research and careful organization or information, this should become possible, because BIRS was designed to be extensible, not only to incorporate new maps or other forms of data representation, but also to incorporate additional or new forms of search engines, thereby providing a useful one-stop tool for information retrieval sessions. All this, together with the suggestions from the subjects in the user study, provides much scope for future work on BIRS.

## References

Bates, M.J. (1986). Subject access in online catalogs: Design model. Journal of the American Society for Information Science, 37(6), 357–376.

Bates, M.J. (1998). Indexing and access for digital libraries and the Internet: Human, database, and domain factors. Journal of the American Society for Information Science, 49(13), 1185–1205.

Bollacker, K.D., Lawrence, S., & Giles, C.L. (1998). CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In. K.P. Sycara & M. Wooldridge (Eds.), Proceedings of the 2nd international conference on autonomous agents (pp. 116–123). New York: ACM Press.

Braam, R.R., Moed, H.F., & van Raan, A.F.J. (1991). Mapping of science by combined co-citation and word analysis. I: Structural aspects. Journal of the American Society for Information Science, 42(4), 233–251.

Bradley, P. (1995). Towards a common user interface. Aslib Proceedings, 47(7/8), 179–184.

Byrne, C.C., & McCracken, S.A. (1999). An adaptive thesaurus employing semantic distance, relational inheritance and nominal compound interpretation for linguistic support of information retrieval. Journal of Information Science, 25(2), 113–131.

Chalmers, M. (1994). Ongoing work on bead: Exploration in information visualization. In: R. Leon (Ed.), Information retrieval: New system and current research (pp. 53–66). London: Talor Graham.

Chen, C. (1999). Visualising semantic spaces and author cocitation networks in digital libraries. Information Processing & Management, 35, 401–420.

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. Information Processing and Management, 27(5), 405–432.

Chen, H., & Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man, and Cybernetics, 22(5), 885–902.

Chen, H., & Ng, T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound search vs. connectionist hopfield net activation. Journal of the American Society for Information Science, 46(5), 348–369.

Chen, H., Yim, T., Fye, D., & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. Journal of the American Society for Information Science, 46(3), 175–193.

Chen, H., Ng, T.D., Martinez, J., & Schatz, B.R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. Journal of the American Society for Information Science, 48(1), 17–31.

Chen, H., Houston, A. L., Sewell, R.R., & Schatz, B.R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. Journal of the American Society for Information Science, 49(7), 582–603.

Chen, H., Martinez, J., Kirchhoff, A., Ng, T.D., & Schatz, B.R. (1998). Alleviating search uncertainty through concept associations: Automatic indexing, co-occurrence analysis, and parallel computing. Journal of the American Society for Information Science, 49(3), 206–216.

Chowdhury, G.G. (1999). The Internet and information retrieval research: A brief review. Journal of Documentation, 55(2), 209–225.

Chowdhury, G.G., & Chowdhury, S. (1999). Digital library research: major issues and trends. Journal of Documentation, 55(4), 409–448.

Croft, W.B. (1995). What do people want from information retrieval? D-Lib Magazine, November. Available: htt;://www.dlib.org/dlib/november95/11croft.html.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. Ergonomics, 35(10), 1297–1326.

Ding, Y. (1998a). Bibliometric analysis of information transfer in the field of information retrieval. First Year Report, Information Studies Division, School of Applied Science, Nanyang Technological University, Singapore.

Ding, Y. (1998b). Visualization of intellectual structure in information retrieval: Author cocitation analysis. International Forum on Information and Documentation, 23(1), 25–36.

Ding, Y., Chowdhury, G.G., & Foo, S. (1999a). Mapping intellectual structure of information retrieval: An author cocitation analysis, 1987–1997. Journal of Information Science, 25(1), 67–78.

Ding, Y.; Chowdhury, G.G., & Foo, S. (1999b). Mapping the development in information retrieval specialty: A bibliometric analysis via journals. In C.A. Macias-Chapula (Ed.), Seventh conference of the international society for scientometrics and informetrics proceedings—1999 (pp. 139–149). Mexico, July 5–9, 1999.

Ding, Y., Chowdhury, G.G., & Foo, S. (2000a). Journal as markers of intellectual space: Journal co-citation analysis of Information Retrieval area, 1987–1997. Scientometrics, 47(1), 55–73.

Ding, Y., Chowdhury, G.G., & Foo, S. (2000b). Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing and Management, submitted.

Doreian, P. (1988). Testing structural-equivalence hypotheses in a network of geographical journals. Journal of the American Society for Information Science, 39(2), 79–85.

Duplenko, Y.K., & Burchinsky, S.G. (1995). Computer-aided clustering of citation networks as a tool of mapping of research trends in biomedicine. Scientometrics, 32(3), 247–258.

Fox, K.L., Frieder, O., Knepper, M.M., & Snowberg, E.J. (1999). SENTINEL: A multiple engine information retrieval and visualization system. Journal of the American Society for Information Science, 50(7), 616–625.

Frants, V.I., Shapiro, J., Taksa, I., & Voiskunskii, V.G. (1999). Boolean search: Current state and perspectives. Journal of the American Society for Information Science, 50(1), 86–95.

Furnas, G.W., et al. (1982). Statistical semantics: How can a computer use what people name things to guess what things people mean when they name things? In Proceedings of the human factors in computer systems conference (pp. 251–253), March 15–17, 1982, Gaitherburg, MD. New York: Association for Computing Machinery.

Gaines, B.R.; Chen, L.L., & Shaw, M.L.G. (1997). Modeling the human factors of scholarly communities supported through the Internet and World Wide Web. Journal of the American Society for Information Science, 48(11), 987–1003.

Garfield, E.; Kimberley, R., & Pendlebury, D.A. (1988). Mapping the social sciences: The contribution of technology to information retrieval. Essays of an information scientist: Science literacy, policy, evaluation, and other essays, Philadelphia: ISI Press, 11, 369–374.

Gomez, L.M., Lochbaum, C.C., & Landauer, T.K. (1990). All the right words: finding what you want as a function of richness of indexing vocabulary. Journal of the American Society for Information Sience, 41(8), 547–559.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. Information Processing & Management, 35, 141–180.

Grivel, L., Polanco, X., & Kaplan, A. (1997). A computer system for big scientometrics at the age of the World Wide Web. Scientometrics, 40(3), 493–506.

Harter, S.P., & Cheng, Y. (1996). Colinked descriptors: Improving vocabulary selection for end-user searching. Journal of the American Society for Information Science, 47(4), 311–325.

Hearst, M.A. (1995). TileBars: Visualization of term distribution information in full text information access. In I. Katz et al. (Eds.), Mosaic of creativity: CHI'95: Proceedings of the ACM/SIGCHI conference on human factors in computing systems; 1995, May 7–11; Denver, CO.

Hetzler, B., Harris, W.M., Havre, S., & Whitney, P. (1998). Visualizing the full specture of document relationships. InProceedings of the 5th international ISKO conference: Structures and relations in the knowledge organization.

Ingwersen, P. (1984). A cognitive view of three selected online search facilities. Online Review, 8(5), 465–492.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. Journal of Information Science, 13, 261–276.

Kohonen, T. (1998). Self-organization of very large document collections: State of the art. In: J. Niklasson, M.Boden, & T. Ziemke (Eds.), Proceedings of ICANN98, the 8th international conference on artificial neural networks (pp. 65–74). London: Springer.

Kostoff, R.N., Eberhart, H.J., & Toothman, D.R. (1997). Database tomography for information retrieval. Journal of Information Science, 23(4), 301–311.

Lawrence, S., & Giles, C.L. (1998). Searching the World Wide Web. Science, 280(3), 98–100.

Mackinlay, J.D.; Rao, R., & Card, S.K. (1995). An organic user interface for searching citation links. In I. Katz et al. (Eds.), Mosaic of creativity: CHI'95: Proceedings of the ACM/SIGCHI conference on human factors in computing systems; 1995, May 7–11; Denver, CO.

McCain, K.W. (1984). Longitudinal author cocitation mapping: The changing structure of macroeconomics. Journal of the American Society for Information Science, 35(6), 351–359.

McCain, K.W. (1986). Cocited author mapping as a valid representation of intellectual structure. Journal of the American Society for Information Science, 37(3), 111–122.

McCain, K.W. (1990). Mapping authors in intellectual space: A technical overview. Journal of the American Society for Information Science, 41(6), 433–443.

McCain, K.W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. Journal of the American Society for Information Science, 42(4), 290–296.

McCain, K.W. (1998). Neural networks research in context: A longitudinal journal cocitation anlaysis of an emerging interdisciplinary field. Scientometrics, 41(3), 389–410.

Nowell, L.T., & Hix, D. (1993). Visualizing search results: User interface development for the project envision database of computer science literature. In M.J. Smith & G. Salvendy (Eds.), Proceedings of the 5th international conference on human-computer interaction; 1993 August 8–13; Orlando, FL.

Nowell, L.T., France, R.K.; Hix, D.; Heath, L., & Fox, E.A. (1996). Visualizing search results: Some alternatives to query-document similarity. In H.P. Frei, D. Harman, P. Schauble, & R. Wilkinson (Eds.), SIGIR'96: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 19th Annual International Conference on Research and Development in Information Retrieval (pp. 67–75). August 18–22; Zurich, Switzerland. New York, NY: ACM.

Noyons, E.C.M., Moed, H.F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. Journal of the American Society for Information Science, 50(2), 115–131.

Peat, H.J., & Willett, P. (1991). The limitation of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science, 42(5), 378–383.

Powers, S. (1998). Dynamic Web Publishing (2nd ed.). Indianapolis, Ind. Sams.net Publishing.

Price, D.J. De Dolla. (1966). The science of scientists. Medical Opinion and Review, 1(10), 88–97.

Quoniam, L., Balme, F., Rostaing, H., Giraud, E., & Dou, J.M. (1998). Bibliometric law used for information retrieval. Scientometrics, 41(1–2), 83–91.

Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. Journal of the American Society for Information Science, 50(8), 639–651.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Saracevic, T., Mokros, H., Su, L.T., & Spink, A. (1991). Interaction between users and intermediaries in online searching. In Proceedings of the 12th annual national online meeting (pp. 229–341).

Schatz, B.R., Johnson, E.H., Cochrane, P.A., & Chen, H. (1996). Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. Digital Library Initiative, Grainger Engineering Library Information Center, University of Illionis at Urbana–Champaign.

Seow, H.S. (1999). Characteristic and user interaction with web search engines: A study on engineering students. Master Thesis. Information Studies Division, School of Applied Science, Nanyang Technological University, Singapore.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265–269.

Small, H., & Griffith, B.C. (1974). The structure of scientific literatures. I. Identifying and graphing specialties. Science Studies, 4, 147–159.

Small, H. (1994). A SCI-MAP case study: Building a map of AIDS research. Scientometrics, 30(1), 229–241.

Small, H. (1997). Update on science mapping: Creating large document spaces. Scientometrics, 38(2), 275–293.

Small, H. (1998). A general framework for general large-scale maps of science in two or three dimensions: The SciViz system. Scientometrics, 41(1–2), 125–133.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50(9), 799–813.

Tijssen, R.J.W., & van Raan, A.F.J. (1990). Net citation balances: A measure of influence between scientific journals. Journal of the American Society for Information Science, 41(4), 298–304.

Trivison, D. (1987). Term co-occurrence in cited/citing journal articles as a measure of document similarity. Information Processing & Management, 23(3), 183–194.

Van Raan, A.F.J. (1997). Scientometrics: State-of-the-art. Scientometrics, 38(1), 205–218.

Van Rijsbergen, C.J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation, 33(2), 106–119.

Voorbij, H.J. (1999). Searching scientific information on the Internet: A Dutch academic user survey. Journal of the American Society for Information Science, 50(7), 598–615.

White, H.D., & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 32, 163–171.

White, H.D., & McCain, K.W. (1997). Visualization of literatures. Annual Review of Information Science and Technology, 32, 99–168.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), 327–355.

Wise, J.A., et al. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In: N. Gershon & S. Eick (Eds.), Proceedings: Information visualization; 1995, October 30–31; Atlanta, GA.

Wise, J.A. (1999). The ecological approach to text visualization. Journal of the American Society for Information Science, 50(9), 814–835.

Woodward, J. (1996). Cataloguing and classifying information resources on the Internet. Annual Review of Information Science and Technology, 31, 189–219.

Yermish, I.R.A. (1975). A citation based interactive associative information retrieval system. Philadelphia, PA: University of Pennsylvania (Ph.D. dissertation). Available from: UMI, Ann Arbor, MI.