J. T. Yao, Y.Y., Yao, "Web-based Information Retrieval Support Systems: building research tools for scientists in the new information age ", Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), Halifax, Canada, Oct 13-17, 2003, pp570-573.

Web-based Information Retrieval Support Systems: building research tools for scientists in the new information age

J.T. Yao Y.Y. Yao
Department of Computer Science, University of Regina
Regina, Saskatchewan, CANADA S4S 0A2
E-mail: {jtyao, yyao}@cs.uregina.ca

Abstract

The concept of Web-based Information Retrieval Support Systems (WIRSS) is introduced. The needs for WIRSS are shown by a detailed case study of existing research article indexing and citation analysis systems, such as Current Content, DBLP, Science Citation Index and CiteSeer. The objective of WIRSS is to build new and effective research tools for scientists to access, explore and use information on the Web, which may lead to improved research productivity and quality.

1 Introduction

An important activity of scientists is to keep updated with current research in the field. A scientist uses many means of communication, tools, and services to find relevant and useful articles [5, 12]. Library, a collection of recorded human knowledge, is a major information resource for scientists. Librarians serve as mediators between human and library collections in order to maximize the utilization of records for the benefit of society [9]. The organization, the process and functionality of library have changed dramatically with the introduction of computer technology especially the Web [10]. The changes have a great impact on the ways in which scientists conduct research.

The Web provides a new medium for storing, presenting, gathering, sharing, processing and using information. It brings us to a new information age. There is a tremendous amount of online materials, such as newspapers, movies, music, journals, and many other products and services. Conceptually, the Web may be viewed as a large and searchable virtual library [10]. The problem of making effective use of the Web for research is a challenge for every scientist. Scientists face many challenges in using Web-based information resources, such as information overload, misinformation, fees, poorly designed navigation, retrieval, and browsing tools [5]. There is an urgent need for new information systems that support research activities and play the roles of traditional librarians.

Web-based Information Retrieval Support Systems (WIRSS) assist the basic research activities, such as retrieval, exploration, organization, and utilization of information on the Web [14, 15]. The goals of WIRSS research are to build new and more effective research tools and systems for scientists to take full advantages of the Web. With such tools, scientists are able to change the Web into a vast and personalized knowledge base.

Having identified the needs for WIRSS, it is very tempting to create new theories and build new systems, as being commonly done. However, the huge number of implemented systems in comparison with a relatively small number of systems actually used in practice suggests that such a straightforward method may be ill-fated. Therefore, we take an alternative approaches known as case study [13], which unfortunately did not receive too much attention. We analyze a few widely used systems, as well as their evolution with the development of computer technology. Such a case study will help us to understand *why* certain systems work and *how* they evolve and adapt to work better. The results bring more insights into WIRSS, establish a solid basis for the study of WIRSS, and provide guidelines for the implementation of new systems.

In finding useful articles, one can use three types of information, bibliography, citation, and full text. The introduction of computer affects the use of such information. With the Web, one can easily extract and present such information. This leads to the adaptation of existing systems on the Web platform. We study several such systems, including bibliography systems [3, 4] and citation analysis systems [2, 11]. The results show that one needs to continually extend the functionalities of each type of systems and integrate different systems. There is a need for extracting and using more types of information in supporting research. The introduction of WIRSS would provide a unified framework to combine research results from many related fields.

2 Research Articles Indexing

Bibliographic information is the least information used in indexing to represent an article. Indexing was one of the most important duties of librarians in traditional libraries. Similar to the evolution of library, scientific bibliography changed from printed to digital media, and from digital media to the Web. Editions of Current Content (CC) experienced all stages from the printed edition to the Web [6]. In contrast, the later DBLP system works on the Web platform.

Evolution of the Current Content

Institute for Scientific Information (ISI) published the first printed edition of the Current Content in 1958. The book editions contain mainly the table of contents of selected scientific and technical journals. One can browse the printed editions to find interesting articles. Additional indexes such as author names make search fast. Electronic editions, *CC on Diskette* and *CC on CD*, were later introduced. Additional features and functionalities were also included in the CC electronic media. One can browse and search abstracts through author names, journal titles, keywords etc., as well as a variety of Boolean combinations. The Web edition, *Current Content Connection*, also provides search and access functionalities with powerful Web technologies. Some new functionalities that could not be possible implemented without the Web technologies are: daily update; automatic email alert with predefined search criteria; access to full text documents of online journals; and creation and use of personalized search profiles [6].

DBLP

Like CC, the DBLP is another implementation of the same idea to provide bibliographic information of articles on the Web. The DBLP [4] contains bibliographic information of scientific journals and proceedings in some fields of Computer Science. By taking advantages of the new medium, the DBLP links together entities within its databases, provides hyperlinks to other databases and systems, and maintains an author tree. Hyperlinks to author homepages, coauthors, conference proceedings and journals are given. Some articles are provided with an EE link that connects to the abstract or full text of the paper. In summary, a salient feature of DBLP is that it clearly expresses and explores new structures derivable from bibliographic databases. The DBLP is a free service, which is in fact a main feature of the Internet. Making research articles online for free access benefits users in search for information and leads to a greater distribution, and therefore increases the chances of citation [8]. Some information in DBLP is collected and extracted automatically. This leads to the problem of misinformation. For example, there may be multiple entries for a single researcher.

3 Citation of Research Articles

Merely listing bibliographic information of articles is not enough for researchers. Research cannot be started from the air as it is normally based on the results of others. One often cites published work to provide background information and acknowledge others' contributions. It is therefore necessary to study and explore the structures of a document collection based on citation and co-citation. Citation index study collects and analyzes the citation information of articles. The results from citation analysis can be used in several ways. One can search for useful information through citation structures. One can also use citation counts for quality evaluation. The quality of a paper can be determined by the number of papers citing it. Two well known examples of citation analysis systems are ISI's SCI [1] and NEC's CiteSeer [2].

Science Citation Index

Science Citation Index (SCI) is perhaps one of the most used and reliable citation indexing systems. SCI provides access to bibliographic information, author abstracts, and cited references in ISI's databases.

The printed editions of SCI faced the same problems as its sister product CC, namely "data retrieval was tedious and time consuming [11]." The search abilities are limited and not easy to use. In addition, the information is updated less frequently. The drawbacks of printed editions limited its potential impacts. SCI became more popular among libraries after introducing a tape stored edition. The new technologies have moved citation data from printed to electronic format, and ultimately into a Web-based environment of hypernavigation, optimistic and context-sensitive linking, and beyond [11]. The Web creates a powerful search and browsing research environment that helps researchers stay up-to-date in their specialities. The SCI Web edition also provides one source for a variety of research data including author abstracts, author addresses, and more information per bibliographic record then in other resources [1]. Cited reference search is one of important features of SCI. Users can identify more recent articles on the same topic by the way of referencing a given article. This is also an example of document space granulations [14].

CiteSeer

The CiteSeer or ResearchIndex "aims to improve the dissemination and feedback of scientific literature, and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness [2]." It not only provides citation information but also links to cached research articles in different formats that allow quick and easy viewing. Similar to the DBLP, the CiteSeer is a

citation product on the Web. It also provides algorithms, techniques, and software to the public free of charge. We may view the CiteSeer as a portal to these freely available online articles. There are two major search functions of the CiteSeer: search documents and search citation. One is based on the documents indexed by the CiteSeer as in CC using full text. The other is based on the citations made by indexed documents as in SCI. The CiteSeer autonomously creates a citation index that can be used for literature search and evaluation[7]. It also shows the context of citations to a given paper, allowing a researcher to quickly and easily see what other researchers have to say about an article of interest. Another feature of the CiteSeer is the awareness and tracking feature that is very useful for researchers to located most recent articles with one's interests. It provides automatic notification of new citations to a given paper, and new papers matching a user profile.

The CiteSeer has other facilities derived from full text analysis. Related documents analysis locates related documents using citation and word based measures and displays an active and continuously updated bibliography for each document. Similar documents analysis shows the percentage of matching sentences between documents. Query-sensitive summaries provide the context of how query terms are used in articles instead of a generic summary, improving the efficiency of search. Citation graph analysis produces visual representation of citations. The CiteSeer also provides statistics of most accessed documents.

The CiteSeer is a good example of applying many techniques, such as citation analysis, text analysis and Web log analysis into one integrated system. One is able to view and compare different structures. For example, we can compare the structures obtained from the above three analyses to gain understanding of the document collection.

4 Web-based Information Retrieval Support Systems

From the studies on various indexing and citation products, we summarize the features of different editions in Table 1. This lays the foundation of the design of information retrieval support systems in line with the study of information retrieval products and their evolution in the Internet age [15].

The systems discussed have different functionalities and targeted domains. The collection criteria of CC and SCI is limited to high quality journals according to ISI's standard. The DBLP collects available table of contents of journals and conference proceedings for some research areas in computer science. The information collection procedures of CC, DBLP, SCI are manually or semi-manually. The CiteSeer automatically collects articles found by crawling mechanisms and submissions from authors. We summarize the functionalities of the Web editions of the four products in Table 2.

Table 1. A comparison of different editions of bibliography and citation products

	Printed	Digital	Web
Full text	Impossible	Maybe	Hyperlink
Abstract	Not	Yes	Yes
Author info.	Limited	Email	Email, Home page
Search	Not	Yes	Yes
Delivery cost	Expensive	Cheap	Cheap &Free
Reproduce	Costly	Cheap	Cheap & Free
Citation info.	Hard	Medium	Easy
Impact	Low	Medium	High

The analysis of existing systems and products leads to an important conclusion. The needs for the design and implementation of new generation systems that explore additional structures and provide more functionalities are obvious. We suggest the term Web-based information retrieval support systems (WIRSS) for such a study. WIRSS are designed with the objective to provide the necessary utilities, tools, and languages that support a user to perform various tasks in finding useful information and knowledge [15]. They can be designed as an integrated systems combining existing systems. Information retrieval support systems, Web browsers, and Web search engines extend the basic search functionalities of data retrieval systems exemplified by a database system. They provide basic functionalities to assist a user in the context of libraries and in the early stage of the Web. A user may need to perform many different tasks when finding useful information. The new tasks include understanding, analysis, organization, and discovery, in addition to the conventional tasks of search and browsing. WIRSS is actually a natural evolution from information retrieval systems (IRS) [15]. The evolution from data retrieval systems to information retrieval systems and from information retrieval systems to information retrieval support systems were discussed in details in [15].

We can classify WIRSS models into three related types. Documents in a document collection serve as the raw data of WIRSS. The document models deal with representations and interpretations of documents and the document collection. They allow multi-representation of documents. The retrieval models deals with the search functionality. They provide languages and tools to assist a user to performs tasks such as searching and browsing. WIRSS should provide multi-strategy retrieval. A user can choose different retrieval models with respect to different document models. The presentation models deal with the representations and

Table 2. A list of functionalities of CC, DBLP, SCI and CiteSeer

	CC	DBLP	SCI	CiteSeer
Search	Yes	Yes	Yes	Yes
Full text link	Partial	Partial	Partial	Yes
Collection	Journal	Journal & Conference	Journal	Digital files on Web
Citation	No	Partial	Yes	Yes
Extraction Citation	No	No	No	Yes
Access Cost	A fee	Free	A fee	Free
Clustering	Journal issue	Journal, Conference, Author	Article referencing	Article, Similar, Access
Verification	High	Medium	High	Low
Automation	Low	Medium	Low	High

interpretations of results from the search. They allow a user to view and arrange search results from various document models. The same results can be viewed in different ways by using distinct presentation models. Moreover, a user can analyze and compare results from different retrieval models.

A single document model, a retrieval model, or presentation model may not be suitable for different types of users. Therefore, WIRSS must support multi-model, and provide tools for users to manage various models. A WIRSS focuses on the supporting functionalities of information retrieval. In contrast, existing IRS only focus on the search and browsing functionalities. WIRSS are more flexible and combine the functionalities of IRS, Web browser and Web search engines. A WIRSS is based on a different design philosophy that emphasizes the supporting functionality of the system, instead of the specific search and browsing functionalities. In the process of finding useful information, a user plays an active role in a WIRSS by using the utilities, tools, and languages provided by the system. The components of a WIRSS are very similar to decision support systems and intelligent systems such as data management, model management, knowledge-based management, and user interface subsystems.

5 Concluding Remarks

The extension of CC and DBLP to include more structures and functionalities would support research activities. The application and adaptation of existing methodologies, such as citation analysis, text analysis and Web log analysis, on the Web platform will result more effective research tools. Our aim is to design and implement intergraded systems under the umbrella of WIRSS. The results of this preliminary study establishes a solid basis of WIRSS. The results show that one need to continually extend the functionality of each type of systems and integrate different systems. The introduction of WIRSS would provide a unified framework to combine research results from many related fields.

References

[1] H. Atkins, The ISI Web of Science: links and electronic journals, *D-Lib Magazine*, 5(9), 1999.

[2] CiteSeer: http://citeseer.nj.nec.com/cs

[3] CSB (Collection of Computer Science Bibliographies): http://liinwww.ira.uka.de/bibliography/index.html

[4] DBLP: http://www.informatik.uni-trier.de/∼ley/db/

[5] D. B. Hoggan, Challenges, strategies, and tools for research scientists: using Web-based information resources, *Electronic Journal of Academic and Special Librarianship*, 3(3), 2002.

[6] ISI: http://www.isinet.com/isi/

[7] S. Lawrence, C. L. Giles, and K. Bollacker, Digital libraries and autonomous citation indexing, *IEEE Computer*, 32(6), 67-71, 1999

[8] S. Lawrence, Online or invisible?, *Nature*, 411(6837), 521, 2001

[9] W. Marsterson, *Information Technology and the Role of the Librarian*, Croom Helm, 1987.

[10] L. M. Saunders (Ed.), *The Virtual Library: Visions and Realities*, Meckler Publishing, 1993.

[11] H. Szigeti, The ISI Web of $\mathrm{Knowledge}^{\mathrm{sm}}$ platform: current and future directions, 2001, http://www.isinet.com/presentrep/essayspdf/wokplat.pdf

[12] C. Tenopir and D.W. King, What do we know about scientists' use of information?, *Proceedings of the Online Meeting*, London, December 2001.

[13] R.K. Yin, *Case Study Research, Design and Methods*, Newbury Park, Sage Publications, 1994.

[14] J. T. Yao and Y.Y. Yao, Information granulation for Web based information retrieval support systems, *Proceedings of SPIE Vol. 5098*, pp138-146, 2003.

[15] Y.Y. Yao, Information retrieval support systems, *FUZZ-IEEE'02* in *The 2002 IEEE World Congress on Computational Intelligence*, pp773-778, 2002.