

Information Retrieval Support Systems

Y.Y. Yao

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: yyao@cs.uregina.ca

Abstract - Information retrieval support systems (IRSS) 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. While existing information retrieval systems (IRS) focus on the search and browsing functionalities, an IRSS focuses on the supporting functionalities. IRSS are more flexible and combine the functionalities of IRS, Web browser and Web search engines. One objective of the paper is to demonstrate the needs for, and the potential benefits of, moving from IRS to IRSS. On the one hand, IRSS is an emerging important research topic, and on the other hand, there is a lack of a systematic study on the topic. Another objective of the paper is to present a framework for IRSS by drawing results from decision support systems (DSS) and intelligent systems. Basic issues of IRSS are discussed, and basic components of an IRSS, as well as its functionalities, are studied.

I. INTRODUCTION

With the rapid growth of information and easy access of information, in particular the boom of the World Wide Web, the problem of finding useful information and knowledge becomes one of the most important topics in information and computer science [1], [2]. Web browsers, Web search engines designed based on the theory of information retrieval (IR), and information retrieval systems (IRS) are some of the solutions to this problem. They aim at providing a user with useful and relevant information in response to a user query.

IRS, Web browsers, and Web search engines extend the basic search functionalities of data retrieval systems (DRS) exemplified by a database system [1], [10], [11]. A major difference between data retrieval (DR) and IR lies in the nature of their problem domains [10]. DR deals with well defined, structured and simple problems, where data items, queries, and matching methods can be precisely defined and interpreted. In contrast, IR deals with not-so-well defined, semi-structured or unstructured, and more complicated problems, where information items (documents), user information needs and queries, and matching methods can not be precisely de-

finied. In an DRS, a user can only perform a well structured task of search. In other word, a user needs to supply a query and the system provides results based on an exact match of data items and the query. With an IRS, a user can perform less structured tasks. The evolution from DRS to IRS increases the power of a user in finding useful information.

Current IRS, Web browsers, and Web search engines provide basic functionalities to assist a user in the context of libraries and in the early stage of the Web. When finding useful information, a user may need to perform more tasks, such as understanding, analysis, organization, and discovery, in addition to the conventional tasks of search and browsing. With the recent development of XML (eXtensible Markup Language), it is possible to express both the structure and semantics information about a document. A user can perform additional tasks with respect to an XML document collection [13]. It is expected that current IRS need to be extended to support more user tasks. The next evolution of retrieval systems is to move from IRS to information retrieval support systems (IRSS). IRSS 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 an IRSS by using the utilities, tools, and languages provided by the system.

In this paper, we briefly discuss some fundamental issues of IRSS. The objective is to demonstrate that IRSS is the next stage in the evolution of retrieval systems. In other words, retrieval systems evolve from DRS to IRS, and from IRS to IRSS. We discuss the importance, the needs, and the potential impact of the design philosophy of IRSS. It is hoped that the new philosophy and principles will lead to more effective retrieval systems.

II. EVOLUTION OF RETRIEVAL SYSTEMS

In order to appreciate the new design philosophy of IRSS, we examine the evolution process of retrieval systems and the roles played by each type of retrieval systems. IRSS is the next stage in the evolution, which deals with more difficult and more complex problems, and provides more supporting functionalities to a user.

A. From Data Retrieval Systems (DRS) to Information Retrieval Systems (IRS)

In one of the classic books on information retrieval (IR), van Rijsbergen compared data retrieval (DR) and IR to illustrate the range of complexity of each mode of retrieval, as well as their common features and differences [10]. DRS may be considered as an early stage, and IRS as the next evolution stage in the development of retrieval systems.

In both modes of retrieval, on the one hand, there is a set of information items (data or documents), and on the other hand, there is a user information need. The function of a retrieval systems is to match items with the user need. In other words, both DR and IR focus on the retrieval functionality, namely, the match of items and user information needs. It is not surprising that many results from the discipline of pattern recognition, which deals with pattern match, are applied into IR [3].

The differences between DR and IR can be seen from the ways in which information items and user information needs are represented, as well as the matching process [1]. A database system is a typical example of DR. In DR, data items and user information needs can be precisely described by using well understanding knowledge representation schemes and query languages. The model is deterministic in the sense that the relationships between data items and user needs are well and objectively defined, and consequently exact match and deductive inference can be used. In contrast, in IR, documents and user needs can not be precisely described, and their relationships are ill and subjectively defined. The model is non-deterministic in the sense that partial or best match and inductive inference are used. In summary, DR deals with structured, and well-defined problems in which there is no uncertainty, while IR deals with semi-structured, unstructured, and ill-defined problems where uncertainty play a major role.

In the design of IRS, the search functionality of DRS is extended. However, the design philosophy remains to be the same. Both DRS and IRS are designed to provide the search functionality. Some of the reasons for adopting the search functionality centered design philosophy can be seen from the historical roles played by IR systems. IR systems were first introduced in libraries for retrieval information [1]. Baeza-Yates and Ribeiro-Neto divide IR systems into three generations [1]. The first generation is basically an automation of conventional manual catalog search, which allows searches using author name and title. The second generation increases the search functionalities by allowing text based search, such as searching by subject headings, keywords, and more complex queries. The third generation focuses on improved graphical in-

terfaces, hypertext features, and open system architectures. We can conclude that IR systems, in the context of libraries, only attempt to automate the task of search, while other user support functionalities are left to the librarians.

Baeza-Yates and Ribeiro-Neto differentiate two distinct types of user tasks when using a retrieval system, the retrieval (search) task and the browsing task [1]. A retrieval task is normally performed by translating an information need into a query and searching using the query. A browsing task is carried out by looking around in a collection of documents through an interactive interface. During browsing the user information need or objective may not be clearly defined, and can be revised through the interaction with the system. The third generation of IR systems provide more functionalities for browsing. As summarized by Baeza-Yates and Ribeiro-Neto, "Classic information retrieval systems normally allows information or data retrieval. Hypertext systems are usually tuned for providing quick browsing. Modern digital library and Web interfaces might attempt to combine these tasks to provide improved retrieval capabilities."

A careful examination of the three generations of IR systems shows that each later generation provide enhanced or more functionalities of the previous generation. As we move from DR to IR, as well as from one generation of IR systems to another, two dimensions of changes can be observed. One dimension concerns the complexity and the nature of the problems, ranging from structured, semi-structured, to unstructured. DR deals with structured problems, where all involved concepts can be precisely defined. IR deals with semi-structured problem, where some concepts can not be precisely defined. Future systems may deal with unstructured problems. Another dimension concerns the user control or user tasks, ranging from simple to complex. DR systems deal with fact retrieval, IR systems deal with non-fact searching and browsing. Future generations of systems may deal with more complex user tasks, such as analysis, organization, and discovery. This two-dimensional description is adopted from a well known framework for decision support systems (DSS) [9].

B. From Information Retrieval Systems (IRS) to Information Retrieval Support Systems (IRSS)

Many new developments in IR have been made, ranging from multimedia (images, audio or video) retrieval, hypertext retrieval, and digital library to Web information retrieval [1]. With the rapid development of the Web and digital libraries, we have witnessed a wider range of

applications of IR systems, and a renewed interest in IR. In fact, IR systems such as search engines play an important role for the success and boom of the Web. On the other hand, the design philosophy and principles of IR, well discussed in classic textbooks [8], [10], have remained more or the less the same. The observation by Lesk that “the more things change, the more they stay the same” is still applicable today as it was applicable 10 years ago [5]. To a large extent, information retrieval can still be viewed as document retrieval by substituting ‘document’ for ‘information’, as pointed out by van Rijsbergen long time ago [10]. The fundamental ideas of IR systems, namely, indexing and searching, have remained to be the same. IR are perceived as systems that provide the basic search and browsing functionalities.

From the two-dimensional description of DR and IR problems in the last section, we need to consider unstructured problem and more user control in the design and implementation of new information retrieval strategies and systems. We need to have a better understanding of user tasks. The ultimate goal of finding information is to use the relevant information, say, in a decision making process. For example, a researcher may compare, analyze and summarize the relevant information in preparing a scientific article, or evaluating a project proposal. In order to find and extract useful information from a large document collection or Web, as well as effectively use the extracted information in problem solving, a user must play an active role in various tasks, such as browsing, investigating, analyzing, understanding, organizing, and searching the collection. Searching and browsing are only some of the simple and front end tasks. The next generation of IR systems must support more types of user tasks, in addition to searching and browsing. Since these tasks can not necessarily be described precisely, fully automation can not be expected. Instead, one can build various tools, methodologies, and languages for supporting such user tasks.

From the above discussion, we can conclude that the search centered philosophy for the design of IR system may no longer be suitable. This is also evident from the fact that many Web search engine users must spend more time to understand, filter, and organize documents returned by a search engine. With the Web as a new media for information storage, delivery, gathering, sharing, processing, and utilization, the problem of information retrieval is no longer a simple process of search.

A new set of philosophy and principles for the design and implementation of the next generation IR systems is needed. Instead of focusing on the search functionality, one focuses on the supporting of functionality [13]. This can be viewed as the next stage in the evolution

of retrieval systems, which leads to the introduction of information retrieval support systems (IRSS). The objective of an IRSS is to support many different types of user tasks in finding and utilizing information, in a similar way that a decision support system (DSS) assists users in making managerial decisions [9].

C. Uses of the Phrase IRSS

The phrase of “information retrieval support system(s)” have been used by many authors and appeared in many places. By querying the search engine Google¹ with exact phrase in January 2002, we obtained 26 hits. We list sample uses of the term from search results:

- The Information-technology Promotion Agency, Japan (IPA) sponsored a research project entitled “Information Retrieval Support System Based on Information Extraction and Probabilistic Reasoning.” The aim of the project is to “enhance information retrieval processes for user support and troubleshooting.”²
- The Ph.D. program in Information Science & Learning Technologies at University of Missouri lists the “creation of an information retrieval support system” as a sample of learning experiences regarding research initiative.³
- In a FCRP (Family Care Research Program, Michigan State University) Abstract, it is stated that “... patients and families require an extensive knowledge base and an information retrieval support system.”⁴
- Many papers contain the exact phrase “information retrieval support system(s)” [6], [7].

Additional searches are carried out by using parts of “information retrieval support system(s)”. Querying by the exact phrase “retrieval support system(s)” results in 54 hits, and querying by the exact phrase “retrieval support” results in 829 hits.

An analysis of results from Google leads to the following observations. The notion of “information retrieval support” have been considered in many different contexts and applications. There are companies that attempt to provide “information retrieval support”. Many researchers stress the “retrieval support” functionality of the system. On the other hand, the concept of “information retrieval support system(s)” has been used in a narrow sense, which only covers the support of specific user tasks. There is a lack of systematic studies on IRSS, although the phrase has been used. Moreover, meanings

¹<http://www.google.com/>

²<http://www.isl.intec.co.jp/contents/E/proj/BayesIR/>

³http://www.coe.missouri.edu/~sislt/programs/phd_framework.html

⁴<http://www.healthteam.msu.edu/fcrp/pdf/psweb/1-97-7-6.htm>

associated with the phrase are different from, though related to, the intended meaning we want to use in the present study.

III. BASIC ISSUES OF IRSS

In this paper, the term information retrieval is used as a much broader umbrella term to cover any and every user tasks in finding and utilizing information from a collection of documents. Similarly, the term IRSS is used as an umbrella term to describe any and every computerized systems used to support and improve retrieval.

A. The Concept of Information Retrieval Support Systems (IRSS)

Our interpretation of IRSS draws extensively results from the related field of decision support systems (DSS) [9]. While DSS focus on supporting and improving decision making, IRSS focus on supporting and improving retrieval. The philosophy, principles, and techniques from DSS are applicable to IRSS by simply substituting the tasks of "decision making" for the tasks of "information retrieval". This view of IRSS is particular reasonable, if one considers the fact that it is necessary to obtain useful information in order to make intelligent and rational decision.

A classical definition of DSS given by Gorry and Morton [4] defines DSS as "interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems." Turban and Aronson compare different definitions of DSS and suggest the following one [9]:

"Computer-based information systems that combine models and data in an attempt to solve nonstructured problems with extensive user involvement through a friendly user interface."

From the above definition, we want to stress two important features of DSS pertinent to our interpretation of IRSS. One feature is the combination of data and models. Data are raw and uninterpreted facts. In order to make sense of raw data, one needs to construct various models. Therefore, an DSS deals with both data and their interpretations. The other feature is the emphasis on the user involvement. An DSS plays a supporting role in problem solving.

One can give a formal definition of IRSS based on the definition of DSS. However, the definition of IRSS is not the focus of the paper. It is perhaps wise to leave the notion loosely defined until we have gained more insights into the problem. For the time being, we rely on our intuitive understanding of IRSS.

B. Characteristics of IRSS

The problems of finding and using relevant information from a large collection of documents are unstructured problems that can not be easily and precisely described. It is made even more complicated by the fact that a user may not know exactly what is being searched for. In solving the retrieval problems, IRSS are more useful and effective than IRS. Based on the two features of DSS mentioned earlier, we can identify some characteristics that distinguish IRSS from current IRS, and show the advantage of IRSS.

Most current IR systems are based on the principle of indexing and searching. Typically, documents are represented by a set of index terms (keywords) appearing in the documents. A user information need is represented by a query, which is either a Boolean expression or simply a list of index terms. The retrieval process is modeled as a (partial) match between documents and a query. Many retrieval models have been proposed and studied, such as the Boolean model, fuzzy set model, vector space model, and probabilistic model [8], [10], [11].

Several problems can be observed about current IRS. Most systems use a very simple document representation schemes, as well as a single and simple retrieval method. All documents are described in the same level of details. The same document representatives and the same retrieval method are used, independent of users. The structures and semantic information of documents and the document collection are not taken into consideration. Although many studies have addressed some of the problems, such as the use of meta-search engines, construction of user profiles, and clustering of document collection, there is not a unified framework that addresses all of them.

The problems of current IRS stem from their emphasis on the storage and search functionalities, which leads to a lack of consideration of the two important issues, namely, models and user involvement. In other words, an IRS performs search at the raw data level, instead of the model level, and without user interaction. Although recent IRS systems exemplified by Web search engines build hierarchical model to provide semantics interpretation of documents in a collection, the end users are not involved in the model construction process. A remedy to this problem is the use of personalized user profile and personalized bookmarks.

IRSS attempt to resolve the problems of IRS by providing more supporting functionalities. An IRSS provides models, languages, utilities, and tools to assist a user in investigating, analyzing, understanding, and organizing a document collection and search results. These tools allow the user to explore both semantic and structural

information of each individual document, as well as the entire collection.

Three related types of models need to be considered in IRSS. Documents in a document collection serve as the raw data of IRSS. The document models deal with representations and interpretations of documents and the document collection. The retrieval models deal with the search. The presentation models deal with the representation and interpretations of results from the search. A single document model, a retrieval model, or presentation model may not be suitable for different types of users. Therefore, IRSS must support multi-model, and provide tools for users to manage various models.

The document models allow multi-representation of documents. Granular computing (GrC), a methodology for computing at different levels of abstractions, will play an important role in the construction of document models [12], [13], [14]. One can use a hierarchical granulation of document collection, namely, a layered and multi-resolution representation of documents. The same document is represented in less detail at a higher level than at a lower level. With the multi-model capability, a user can create different logical views of a document collection and logical links between documents, as well as compare and investigate various views. The resulting in-depth knowledge can help the user to locate and infer useful information. The recent development of XML enables us to describe both the structural and semantic information of a document. Such information makes the construction of multi-document models an achievable goal.

The retrieval models provide languages and tools to assist a user to perform tasks such as searching and browsing. IRSS should provide multi-strategy retrieval. A user can choose different retrieval models with respect to different document models. The presentation models allow a user to view and arrange search results, as well as 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.

IRSS implement and manage three types of models, as well as the associated languages, tools and utilities. An IRSS is highly interactive so that a user can make decisions at various stages. A user plays a more active role in the process of finding useful information. There are many advantages for the extensive user involvement. The user involvement is particularly important in exploratory type of searching and browsing. The usefulness or the relevance of each information item (i.e., a group of documents, a document, and parts of a document) can only be determined by the user. Without the involvement

of a user in the analysis and organization of a document collection, one can not expect the user to provide a meaningful query.

From the previous discussion, we can conclude that an IRSS performs easy and loosely structured retrieval tasks, where automation is possible and user involvement is not necessary. On the other hand, an IRSS supports a user to perform difficult and unstructured retrieval problems. Since a fully automation is impossible, at least not for the time being, the main function of IRSS is to support a user. The retrieval process is controlled by the user.

C. Components of IRSS

One may argue about the exact components of an IRSS. We take a simple approach by adopting the results from DSS and intelligent systems.

According to Turban and Aronson, a DSS normally consists of four subsystems: [9].

- *Data management subsystem*: This subsystem deals with lower level raw data management using software systems such as database management system (DBMS) and data warehouse.
- *Model management subsystem*: This subsystem is referred to as a model base management system (MBMS). It includes existing quantitative models for analyzing and interpreting the raw data, and provides language and tools for building user models.
- *Knowledge-based management subsystem*: This subsystem supports other subsystem and provides intelligence to a decision maker.
- *User interface subsystem*: This subsystem handles the interaction between user and the system.

The above schematic description of DSS can be applied to the study of IRSS. In other words, although the objects managed by each subsystem may be different, the fundamental principles are the same.

D. Fields Related to IRSS

Techniques, results and lessons from many fields can be used in the study of IRSS and to enhance the capabilities of an IRSS. A few related fields are summarized below, in addition to the previously discussed DSS and GrC.

Expert systems (ES).

A well established practice in expert systems is the separation of knowledge and inference engine. While the inference engine is logic based and problem independent, the knowledge base is domain specific. Expert system shells that implement inference engine can be used to build many different domain-specific expert systems.

Similarly, we can build knowledge based IRSS by separating knowledge base and management subsystem. An IRSS shell can be built that provide a set of domain and user independent tools, using which domain specific IRSS can be constructed.

Users of retrieval systems may fall into many different categories, have different background, and with different types of information needs. Documents in a collection may also cover different domains. One can not expect to design a system that is best for everyone and for every domain. Consequently, one needs to study principles, methodologies, and techniques that can be used to design and implement domain-specific IRSS.

Another feature of expert systems is the explanation functionality. An expert system not only provides a solution, but can also explain why and how the solution is arrived. It is reasonable to insist on the explanation functionality of an IRSS.

Machine learning, data mining and text mining.

By applying algorithms of machine learning, data mining, and text mining to documents stored in an IRSS, one may discover patterns and extract knowledge useful to a user. Such functionalities is particular useful to users who are interested in exploratory searching and browsing. For instance, a user can track trends in a particular area or discover emerging topics from the constantly changing document collection. A user may also discover links between different documents or research areas.

Computer graphics and data visualization.

In many cases, a user may not want details about particular documents that contain the useful information. A user may want to have a general feeling before going to a more in-depth analysis. With the granulation of document collection, it is possible to provide a user with granulated view, in which details are omitted. Most current IRS present search results in the form of ranked list of individual documents. In an IRSS, a user should be able to use graphics and visualization tools to view a particular document model. Visualization enables a user to perform high level inference and analysis.

Intelligent information agents.

Intelligent information agents have been used by many IRS to collect information and interact with users. The potentials of agents need to be further explored in IRSS. In particular, a user should be allowed to construct a personalized agent to interact with an IRSS. The autonomy and learning capabilities of agents make them attractive to both IRSS and users.

In summary, an IRSS can incorporate any particular type of information systems to provide a specific type of support.

IV. CONCLUSION

We introduce the concept of information retrieval support systems (IRSS) and examine perspectives of IRSS. IRSS may be viewed as the next stage in the evolution of retrieval systems. IRSS focus on various supporting functionalities of retrieval systems, in addition to support the search and browsing. IRSS also stress the importance of extensive user involvement.

Our formulation and understanding of IRSS draw extensively results from DSS, which in turn draw results from many fields. The main contribution of the paper is the introduction of a new design philosophy to information retrieval. It fundamentally changes the design goal of current IR systems, which enables us to moves from simple retrieval systems to advanced retrieval support systems.

References

- [1] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*, Addison Wesley, New York, 1999.
- [2] Dominich, S. *Mathematical Foundations of Information Retrieval*, Kluwer Academic Publishers, Dordrecht, 2001.
- [3] Duda, R.O. and Hart, P.E. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
- [4] Gorry, G.A. and Morton, M.S.S. A framework for management information systems, *Soloan Management Review*, **13**, 55-70, 1971.
- [5] Lesk, M. SIGIR'91: the more things change, the more they stay the same, *SIGIR Forum*, **25**, 4-7, 1991.
- [6] Morimoto, Y., Nakayama, M. and Shimizu, Y. A development of information retrieval support system for learning material on the Web, *Transactions of Japanese for Information and Systems in Education*, **17**, 231-240, 2001.
- [7] Sakai, H., Ohtake, K. and Masuyama, S. A retrieval support system by suggesting terms to a user, *Proceedings 2001 International Conference on Chinese Language Computing*, pp. 77-80, 2001.
- [8] Salton, G., and McGill, M. *Introduction to Modern Information Retrieval*, McGraw Hill, New York, 1983.
- [9] Turban, E. and Aronson, J.E. *Decision Support Systems and Intelligent System*, Prentice Hall, New Jersey, 2001.
- [10] van Rijsbergen, C.J. *Information Retrieval*, Butterworths, London, 1979.
- [11] Wong, S.K.M. and Yao, Y.Y. On modeling information retrieval with probabilistic inference, *ACM Transactions on Information Systems*, **13**, 38-68, 1995.
- [12] Wong, S.K.M., Yao, Y.Y., and Butz, C.J. Granular information retrieval, in: *In Soft Computing in Information Retrieval: Techniques and Applications*, Crestani, F. and Pasi, G. (Eds.), Springer Verlag, pp. 317-331, 2000.
- [13] Yao, Y.Y., Song, K. and Saxton, L.V. Granular computing for the organization and retrieval of scientific XML documents, to appear in *Proceedings of the Sixth International Conference on Computer Science and Informatics*, 2002.
- [14] Zadeh, L.A. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, **19**, 111-127, 1997.