Yiyu Yao, Yi Zeng, Ning Zhong, Xiangji Huang. Knowledge Retrieval (KR), In: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Silicon Valley, USA, November 2-5, 2007, 729-735.

Knowledge Retrieval (KR)

Yiyu Yao,[1, 2] Yi Zeng,[2] Ning Zhong,[2, 3] and Xiangji Huang[4]

[1]Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
yyao@cs.uregina.ca

[2]International WIC Institute
Beijing University of Technology
Beijing, P.R. China 100022
yzeng@emails.bjut.edu.cn

[3]Department of Information Engineering
Maebashi Institute of Technology
Maebashi-City, Japan 371-0816
zhong@maebashi-it.ac.jp

[4]School of Information Technology
York University
Toronto, Ontario, Canada M3J 1P3
jhuang@cs.yorku.ca

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

— T.S. Eliot, The Rock, 1934.

Abstract

With the ever-increasing growth of data and information, finding the right knowledge becomes a real challenge and an urgent task. Traditional data and information retrieval systems that support the current web are no longer adequate for knowledge seeking tasks. Knowledge retrieval systems will be the next generation of retrieval system serving those purposes. Basic issues of knowledge retrieval systems are examined and a conceptual framework of such systems is proposed. Theories and Technologies such as theory of knowledge, machine learning and knowledge discovery, psychology, logic and inference, linguistics, etc. are briefly mentioned for the implementation of knowledge retrieval systems. Two applications of knowledge retrieval in rough sets and biomedical domains are presented.

1. Introduction

Although the quest for knowledge from data and information is an old problem, it is perhaps more relevant today than ever before. In the last few decades, we have seen an unprecedented growth rate of data and information. It is necessary to reconsider the question: "Where is the knowledge we have lost in information?"

The data-information-knowledge-wisdom hierarchy is used in information sciences to describe different levels of abstraction in human centered information processing. Computer systems can be designed for the management of each of them. Data Retrieval Systems (DRS), such as database management systems, are well suitable for the storage and retrieval of structured data. Information Retrieval Systems (IRS), such as web search engines, are very effective in finding the relevant documents or web pages that contain the information required by a user. What lacks in those systems is the management at the knowledge level. A user must read and analyze the relevant documents in order to extract the useful knowledge. In this paper, we propose that Knowledge Retrieval Systems (KRS) is the next generation retrieval systems for supporting knowledge discovery, organization, storage, and retrieval. Such systems will be used by advanced and expert users to tackle the challenging problem of knowledge seeking.

While the growth and evolution of the Web makes knowledge retrieval systems a necessity for supporting the future generations of the Web, the extensive results from machine learning, knowledge discovery, in particular, text mining, and knowledge based systems make the implementation of such systems feasible.

Many proposals and research efforts have been made regarding knowledge retrieval [8, 13, 17, 18, 26, 29, 34]. They cover various specific aspects and provide us insights into further development of knowledge retrieval. Those efforts suggest that it is the time to study knowledge retrieval on a grand scale.

In this paper, we argue that knowledge retrieval systems are the natural next step in the evolution of retrieval systems. Specifically, we examine the characteristics and main features of data retrieval systems, information retrieval systems

and knowledge retrieval systems. A conceptual framework of knowledge retrieval system is outlined. The main components in this framework are the discovery of knowledge, the construction of knowledge structures, and the inference of required knowledge based on the knowledge structures. One the one hand, results from existing studies are drawn for the study of knowledge retrieval. On the other hand, knowledge retrieval is studied in its own right by focusing on its unique methodologies and theories. A success of knowledge retrieval will have a great impact on future retrieval systems and future generations of the Web.

2. Data and Information Retrieval vs. Knowledge Retrieval

Knowledge retrieval systems may be considered as the next generation in the evolution of retrieval systems. Their unique features and characteristics become clear by a comparison with existing systems.

2.1 Generations of Retrieval Systems

In information and management sciences, one considers the following hierarchy [23]:

- Data,
- Information,
- Knowledge,
- Wisdom.

It concisely summarizes different types of resources that we can use for problem solving. The hierarchy represents increasing levels of complexity that require increasing levels of understanding [30]. The generations of the retrieval systems may be studied based on this hierarchy. For example, Yao suggests an evolution process of retrieval systems from data retrieval to knowledge retrieval, and from information retrieval to information retrieval support (a step towards knowledge retrieval) [29]. LaBrie argues the needs for a change from data and information retrieval to knowledge retrieval, and considers the problem of retrieving and searching for knowledge objects in advanced knowledge management systems [22].

In the data level, we use data retrieval systems to find relevant data and acquire knowledge. The problem to be solved is well structured, and the concept definitions are clear. Data mining could help us get interesting knowledge from a large volume of data stored in the database.

In the information level, we use information retrieval systems to acquire knowledge. The problem is semi-structured, and the concept definitions are not always clear. We retrieve the relevant information by keywords or their combinations and get the implicit knowledge from retrieved results. This is effective if the scale of the information collection is small. However keyword based matching and page ranking cannot satisfy users needs in the modern Internet age. On the one hand, retrieved results containing the keywords might be huge. Even through page ranking, the most relevant results can be of hundred pages, which will be very difficult for a user to explore one by one. On the other hand, results recommended by page rank might not satisfy various user needs. Some of the most relevant results to specific users might not be ranked in the front portion of the list.

In practical situations, we need to perform more tasks in addition to simple search. Taking scientific literature search as an example, most tasks are not searching for specific articles. We would want to find out which problems have been solved, which ones have no solutions, which fields have more audiences and which topics are promising. It will not be easy to extract such knowledge from increasing volume of information by using current information retrieval systems. This may stem from a discrepancy between knowledge representation methods in retrieval systems and human thoughts.

The storage methods of information in the Web, literature databases, digital libraries are documents, information flows, etc. They are not closely related to the ways that human organize knowledge. People need to find, learn, and reorganize retrieved results to extract and construct knowledge embedded in information.

Those problems cannot be solved satisfactorily in the information level. The task of finding knowledge from information and organizing it into the structures that human can use are the focus of knowledge retrieval systems. The process of retrieving information is changed to retrieving knowledge directly, and the retrieved results is changed from relevant documents to knowledge. Some of the hard problems in information retrieval may find their solutions in knowledge retrieval.

2.2 A Comparison of Retrieval Systems

Knowledge retrieval (KR) focuses on the knowledge level. We need to examine how to extract, represent, and use the knowledge in data and information [2]. Knowledge retrieval systems provide knowledge to users in a structured way. They are different from data retrieval systems and information retrieval systems in inference models, retrieval methods, result organization, etc. Table 1, extending van Rijsbergen's comparison of the difference between data retrieval and information retrieval [27], summarizes the main characteristics of data retrieval, information retrieval, and knowledge retrieval [33].

The core of data retrieval and information retrieval are retrieval subsystems. Data retrieval gets results through Boolean match [1]. Information retrieval uses partial match and best match. Knowledge retrieval is also based on partial

	Data Retrieval	Information Retrieval	Knowledge Retrieval
Match	Boolean match	partial match, best match	partial match, best match
Inference	deductive inference	inductive inference	deductive inference, inductive inference, associative reasoning, analogical reasoning
Model	deterministic model	statistical and probabilistic model	semantic model + inference model
Query	artificial language	natural language	knowledge structure + natural language
Organization	table, index	table, index	knowledge unit and knowledge structure
Representation	number, rule	natural language markup language	concept graph, predicate logic, production rule, frame, semantic network, ontology
Storage	database	document collections	knowledge base
Retrieved Results	data set	sections or documents	a set of knowledge unit

Table 1. A Comparison of Data Retrieval, Information Retrieval, and Knowledge Retrieval

match and best match.

Considering inference perspective, data retrieval uses deductive inference, and information retrieval uses inductive inference [27]. Considering the limitations from the assumptions of different logics, traditional logic systems (e.g., Horn subset of first order logic) cannot make efficient reasoning in a reasonable time [7]. Associative reasoning, analogical reasoning and the idea of unifying reasoning and search may be effective methods of reasoning at the web scale [4, 7].

From retrieval model perspective, knowledge retrieval systems focus on the semantics and knowledge organization. Data retrieval and information retrieval organize the data and documents by indexing, while knowledge retrieval organizes knowledge by connections among knowledge units and the knowledge structures.

3. A Conceptual Framework of KR Systems

3.1 Challenges to Traditional KR Systems

The term "knowledge retrieval" is not new and has been used by many authors. Some authors considered knowledge retrieval as a process of information retrieval [34]. On the other hand, there are authors who investigated this topic in its own right. Frisch discussed knowledge retrieval based on a knowledge base (KB), and considered the entire retrieval process as a form of inference [8]. Oertel and Amir presented an approach to retrieve commonsense knowledge for autonomous decision making [18]. Martin and Eklund examined different metadata languages for knowledge representation (e.g., RDF, OML) and proposed to use general and intuitive knowledge representation languages, rather than XML-based languages to represent knowledge. They proposed methods which satisfy users requirement at different levels of details. Kame and Quintana proposed a concept graph based knowledge retrieval system [13]. In their system, sentences are converted into concept graphs. Chen and Hsiang presented a logical framework of knowledge retrieval by considering fuzziness in inference [5]. To some extent, the framework is restricted to information retrieval and question answering systems. Models and methods for text based knowledge retrieval have also been investigated [17, 26, 34].

In the web age, the traditional understanding of knowledge retrieval faces new challenges. New methodologies and techniques need to be explored.

(1) Traditional knowledge representation methods are investigated in a small scale environment. How to represent knowledge in a large scale needs to be explored. Visualized structured knowledge may be one of the possible strategy.

(2) Traditional knowledge retrieval concentrates on knowledge storage, which is unfortunately not from a human oriented perspective: how to provide and use knowledge in more convenient ways.

(3) Inductive reasoning and deductive reasoning are applicable in data retrieval and small-scale information retrieval. Finding knowledge in the web context needs more effective reasoning methodologies.

(4) Knowledge based systems usually provide knowledge in the form of text. A user needs to extract different views. In knowledge retrieval, knowledge should be examined from different perspectives.

(5) Traditional knowledge based systems assume that the stored contents are all trustworthy, but in the web environment, this assumption is not always valid, as many knowledge are not reliable. Knowledge validation in the complex environment is a challenge.

(6) To some extend, knowledge stored in many systems are static, but in the web environment, knowledge is dynam-

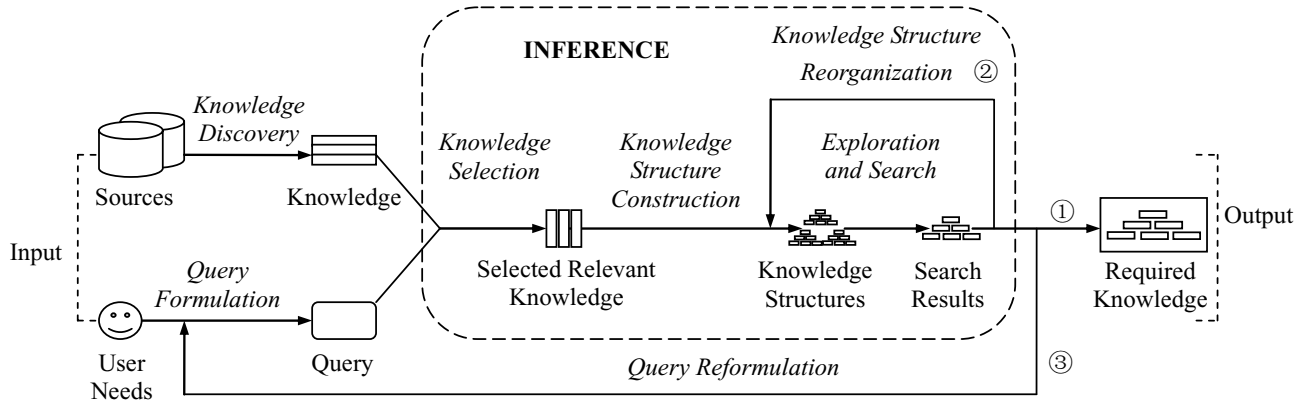

Figure 1. A Typical KR System

ically changing all the time. How to update knowledge in this environment is also a challenge.

3.2 A Typical KR System

Knowledge represented in a structured way is consistent with human thoughts and is easily understandable [20, 26]. Sometimes, users do not know exactly what they want or are lack of contextual awareness [28]. If knowledge can be provided visually in a structured way, it will be very useful for users to explore and refine the query [22].

Figure 1 shows a conceptual framework of a typical knowledge retrieval system. The main process can be described as follows:
(1) **Knowledge Discovery**: Discovering knowledge from sources by data mining, machine learning, knowledge acquisition and other methods.
(2) **Query Formulation**: Formulating queries from user needs by user inputs. The inputs can be in natural languages and artificial languages.
(3) **Knowledge Selection**: Selecting the range of possible related knowledge based on user query and knowledge discovered from data/information sources.
(4) **Knowledge Structure Construction**: Reasoning according to different views of knowledge, domain knowledge, user background, etc. in order to form knowledge structures. Domain knowledge can be provided by expert systems. User background and preference can be provided by user logs.
(5) **Exploration and Search**: Exploring the knowledge structure to get general awareness and refine the search. Through understanding the relevant knowledge structures, users can search into details on what they are interested in to get the required knowledge.
(6) **Knowledge Structure Reorganization**: Reorganizing knowledge structures if users need to explore other views of selected knowledge.
(7) **Query Reformulation**: Reformulating the query if the constructed structures cannot satisfy user needs.

One of the key features of knowledge retrieval systems is that knowledge are visualized in a structured way so that users could get contextual awareness of related knowledge and make further retrieval. Main operations for exploration [12, 28] are browsing, zooming-in, and zooming-out. Browsing helps users to navigate. Zoom-out provides general understanding, while zoom-in presents detailed knowledge and its structure.

We need to point out that the full process contains two levels of feedbacks. One at a local level dealing with knowledge structure reorganization, and the other at a global level dealing with query reformulation. Users' browsing and exploration history will be stored in user logs as background information for improving the personalized knowledge structures. Knowledge retrieval is a typical example of human-centered computing [11], and its evaluation is more related to personal judges, which makes a balance between computer automation and user intervention.

3.3 Knowledge Structures

The definition, representation, generation, exploration and retrieval of knowledge structures are the main issues in knowledge retrieval.

Concept is the basic unit of human thoughts. We can build knowledge based on concepts and the relations among

them [32]. Knowledge structure is built based on hierarchical structures of concepts. In the context of granular computing, a knowledge unit can be considered as a granule. Drawing results from granular structure, knowledge structures can be examined at least at three levels [31]:

– internal structures of knowledge units,
– collective structures of a family of knowledge units,
– hierarchical structures of a web of knowledge units.

A unit of knowledge may be decomposed into a family of smaller units. Their decomposition and relationship represent the internal structures [15]. The collective structures of a family of knowledge units describe the relations of knowledge units in the same level. Different levels of knowledge units form a partial ordering. The hierarchical structures describe the integrated whole of a web of knowledge units from a very high level of abstraction to the very finest details.

Tree structure and concept graph are most commonly used methods for visualizing knowledge structures [14, 16, 26]. Semantic network is used for knowledge representation [20, 24]. Formal concept analysis can be used to generate concept graph [9].

Knowledge unit and knowledge association are the two elements of knowledge [34]. Knowledge unit is considered to be the basic unit of knowledge, and the complex knowledge structures are based on knowledge associations. Knowledge is organized in hierarchies.

Semantic tree is one of the mainly used methods for knowledge representation [20] and the visualized tree structures are convenient for understanding [20, 22]. A semantic tree represents a hierarchical structure [20]. Knowledge can be represented as many semantic trees according to different views and understandings [6, 14, 20]. The knowledge structures discovered and constructed by a knowledge retrieval system should be in multilevels and multiviews.

The discovery of hierarchical structures is to find similar or different features of knowledge and make partitions or coverings for it. Hierarchical clustering is a practical method for knowledge structure construction [25].

The sources of knowledge may have effects on the final knowledge structures. Selecting reliable knowledge from the web environment is one of the key issue for knowledge structure generation. For example, in literature exploration, top journals and conference proceedings may be more important and valuable than some web pages.

We should not generate knowledge structures just from contents stored in files or documents. Various views should be explored. Considering literature on the web, at least content view, structure view and usage view should be explored to generate the multiviews of knowledge structures [32].

When users know exactly what they want, a knowledge retrieval system should provide what they need in a more direct way. The retrieval results should be hierarchical. Coarser results provide general knowledge, while finer results provide detailed knowledge. Users can interact with the system to decide which level of results they want.

4. Theories and Technologies Supporting KR

It is generally believed that new ideas may be repackaging or reinterpretation of old ones [10]. As a new research field, knowledge retrieval can draw results from the following related theories and technologies:

- *Theory of Knowledge*: knowledge acquisition, knowledge organization, knowledge representation, knowledge validation, knowledge management.

- *Machine Learning and Knowledge Discovery*: preprocessing, classification, clustering, prediction, postprocessing, statistical learning theory.

- *Psychology*: cognitive psychology, cognitive informatics, concept formation and learning, decision making, human-computer interaction.

- *Logic and Inference*: propositional logic, predicate logic, attribute logic, universal logic, inductive inference, deductive inference, associative reasoning, analogical reasoning, approximate reasoning.

- *Information Technology*: information theory, information science, information retrieval, database systems, knowledge-based systems, rule-based systems, expert systems, decision support systems, intelligent agent technology.

- *Linguistics:* computational linguistics, natural language understanding, natural language processing.

Topics listed under each entry serve as examples and do not form a complete list.

5. Applications of KR

When scientists explore the literature, they search for the scientific facts and research results. The goal of a literature exploration support system is to provide those in a knowledgable way [32]. Knowledge structures provide a high level understanding of scientific literature and hints regarding what has been done and what needs to be done.

As a demonstration, we examined knowledge structures obtained from an analysis of papers of two rough sets related conferences [32]. Different views provide various unique understanding of the literature. By drawing results from web mining, one examines literature and provides relevant knowledge in multiple views and at multiple levels,

based on the contents, structures, and usages of the literature. The main knowledge structure is constructed based on proceedings indexes, document structures and domain knowledge. The knowledge structures not only provide a relation diagram of specified discipline, but also help researchers to find the contribution of each study and possible future research topics in rough sets [32].

Considering biomedical literature explorations, the relationship and connections of various genes and proteins related publications can be found based on the structure implicit in the genes and proteins. Current biomedical research is characterized by immense volume of data, accompanied by a tremendous increase in the number of gene and protein related publications. However, many interesting links that connect facts, assertions or hypotheses may be missed because these publications are generated by many authors working independently and the functions of many genes and proteins are separately described in the literature.

In order to build knowledge structures for biomedical knowledge retrieval, a large database of abstracts, a subset of PubMed database[1] [19], will be used as our information search space. For example, each gene is mapped to a document roughly discussing the gene's biological function. The literature database is then searched for documents similar to the gene's document. The resulting set of documents typically discusses the gene's function. Since each set corresponds to a gene, the similar document sets can be mapped back to their originating genes in order to establish functional relationships among these genes. The main knowledge structure is constructed based on: first, functional relationships among genes and proteins, on a genome-wide scale; second, the literature specifically relevant to the function of these genes and proteins; third, a short-term list characterizing the document set, which suggests why the genes and proteins are considered relevant to each other, and what their biological functions are.

It is clear that knowledge retrieval systems can greatly enhance our understanding of genetic processes for biomedical research. The associative organization is more closer to human intuition than conventional keyword match [21]. The visualized structures can provide medical doctors and researchers new ideas on medicine use and production.

6. Conclusion

The main contribution for this paper is to suggest knowledge retrieval as a new research field. A comparative study of different levels of retrieval systems is given based on the data-information-knowledge-wisdom hierarchy. The results suggest that different retrieval systems focus on different levels of retrieval problems. A framework of knowledge retrieval system has been proposed. A list of theories and technologies related to KR has been discussed. The applications of KR may show great impact on the future development of retrieval systems, digital libraries, and the Web.

[1] It contains over 17,000,000 scientific abstracts

Acknowledgments

This study was supported in part by research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada. The paper was prepared when Yi Zeng was visiting the University of Regina with financial support from an NSERC discovery grant awarded to Yiyu Yao. The authors would like to thank anonymous reviewers for their constructive suggestions.

References

[1] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval, Addison Wesley, 1999.

[2] Bellinger, G., Castro, D. and Mills, A. Data, Information, Knowledge, and Wisdom, http://www.systems-thinking.org/dikw/dikw.htm (Accessed on August, 16th, 2007).

[3] Benjamins, V.R. and Fensel, D. Community is knowledge! in $(KA)^2$, Proceedings of the 1998 Knowledge Acquisition Workshop, 1998.

[4] Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N. and Weitzner, D.J. A Framework for Web science, Foundations and Trends in Web Science, 2006, 1(1): 1-130.

[5] Chen, B.C. and Hsiang, J. A logic framework of knowledge retrieval with fuzziness, Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, 2004: 524-528.

[6] Collins, A.M. and Quillian, M.R. Retrieval time from semantic memory, Journal of Verbal Learning and Verbal Behavior, 1969, 8: 240-248.

[7] Fensel, D. and van Harmelen, F. Unifying reasoning and search to web scale, IEEE Internet Computing, 2007, 11(2): 96, 94-95.

[8] Frisch, A.M. Knowledge Retrieval as Specialized Inference, Ph.D thesis, University of Rochester, 1986.

[9] Ganter, B. and Wille, R. Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, 1999.

[10] Hawkins, J. and Blakeslee, S. On Intelligence, Henry Holt and Company, 2004.

[11] Jaimes, A., Gatica-Perez, D., Sebe, N. and Huang, T.S. Human-centered computing: toward a human revolution, IEEE Computer, 2007, 40(5): 30-34.

[12] Jain, R. Unified access to universal knowledge: next generation search experience, white paper, 2004.

[13] Kame, M. and Quintana, Y. A graph based knowledge retrieval system, Proceedings of the 1990 IEEE International Conference on Systems, Man and Cybernetics, 1990: 269-275.

[14] Keil, F.C. Concepts, Kinds, and Cognitive Development, MIT Press, 1989.

[15] Langseth, H., Aamodt, A. and Winnem, O.M. Learning retrieval knowledge from data, Proceedings of the 1999 International Joint Conference on Artificial Intelligence Workshop ML-5: Automating the Construction of Case-Based Reasoners, 1999: 77-82.

[16] Martin, P. and Alpay, L. Conceptual structures and structured documents, Proceedings of the 4th International Conference on Conceptual Structures, LNAI 1115, Springer, 1996: 145-159.

[17] Martin, P. and Eklund, P.W. Knowledge retrieval and the World Wide Web, IEEE Intelligent Systems, 2000, 15(3): 18-25.

[18] Oertel, P. and Amir, E. A framework for commonsence knowledge retrieval, Proceedings of the 7th International Symposium on Logic Formalizations of Commonsense Reasoning, 2005.

[19] PubMed database: http://www.pubmed.gov

[20] Quillian, M.R. Semantic Memory, Semantic Information Processing, Minsky, M.L (Eds.), MIT Press, 1968: 216-270.

[21] Robert, H. and Alfonso, V. Implementing the iHOP concept for navigation of biomedical literature, Bioinformatics, 2005, 21: 252-258.

[22] LaBrie, R.C. The Impact of Alternative Search Mechanisms on the Effectiveness of Knowledge Retrieval, Ph.D thesis, Arizona State University, 2004.

[23] Sharma, N. The Origin of the "Data Information Knowledge Wisdom" Hierarchy, http://www-personal.si.umich.edu/ñsharma/dikw_origin.htm (Accessed on August, 16th, 2007).

[24] Sowa, J.F. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole, 2000.

[25] Tenenbaum, J. Knowledge Representation: Spaces, Trees, Features, Course Lecture Notes on Computational Cognitive Science, MIT Open Courseware, Fall 2004.

[26] Travers, M. A visual representation for knowledge structures, Proceedings of the 2nd annual ACM conference on Hypertext and Hypermedia, 1989: 147-158.

[27] van Rijsbergen, C.J. Information Retrieval, Butterworths, 1979.

[28] White, R.W., Kules, B., Drucker, S.M. and Schraefel, M.C. Supporting exploratory search, Communications of the ACM, 2006, 49(4): 37-39.

[29] Yao, Y.Y. Information retrieval support systems, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, 2002, 1092-1097.

[30] Yao, Y.Y. Web Intelligence: new frontiers of exploration, Proceedings of the 2005 International Conference on Active Media Technology, 2005: 3-8.

[31] Yao, Y.Y. Three perspectives of granular computing. Journal of Nanchang Institute of Technology, 2006, 25(2): 16-21.

[32] Yao, Y.Y., Zeng, Y. and Zhong, N. Supporting literature exploration with granular knowledge structure, Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, LNAI 4482, Springer, 2007, 182-189.

[33] Zeng, Y., Yao, Y.Y. and Zhong, N. Granular structure-based knowledge retrieval [In Chinese], Proceedings of the Joint Conference of the Seventh Conference of Rough Set and Soft Computing, the First Forum of Granular Computing, and the First Forum of Web Intelligence, 2007.

[34] Zhou, N., Zhang, Y.F. and Zhang, L.Y. Information Visualization and Knowledge Retrieval [In Chinese], Science Press, 2005.