Information Granulation for Web based Information Retrieval Support Systems

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

In this paper, we discuss the potential applications of data mining techniques for the design of Web based information retrieval support systems (IRSS). In particular, we apply clustering methods for the granulation of different entities involved in IRSS. Two types of granulations, single-level and multi-level granulations, are investigated. Issues of document space granulation, query space granulation, term space granulation, and retrieval results granulation are studied in detail. It is demonstrated that each different granulation supports a different user task.

Keywords: Information retrieval, Granular computing, Information granulation, Information retrieval support systems

1. INTRODUCTION

The Internet and the Web offer new opportunities and challenges to information retrieval researchers. With the information explosion and never ending increase of web pages as well as digital data, it is very hard to retrieval useful and reliable information from the Web. Materials from millions of web pages from organizations, institutions and personnel have been made public electronically accessible to millions of interested users. The Web uses an addressing system called Uniform Resource Locators (URLs) to represent links to documents on web servers. These URLs provide location information. Like titles of books in traditional libraries, no one can remember all URLs on the Web. Web search engines allow us to locate the Internet resources through thousands of Web pages. It is almost impossible to get the right information as there is too much irrelevant and out dated information.

Information retrieval systems provide useful information in libraries to researchers. The Web can be viewed as a virtual library. Information retrieval is an important and major component of the Internet and the Web in the information age and should play an important role in knowledge discovery . General search engines such as, google, AltaVista, Excite are considered as the powerful search engines so far. Most of the current search engines are based on words, not the concepts. When searching for certain information or knowledge with a search engine, one can only uses a few key words to narrow down the search. The result of the search is tens or maybe hundreds of relevant and irrelevant links to various Web pages. We will discuss the potential applications of data mining techniques for the design of Web based information retrieval support systems (IRSS) in this paper.

Granular computing (GrC) is a newly developed technique which has been drawn attention by researchers. It is an umbrella term to cover any theories, methodologies, tools and techniques that make use of granules in problem solving.[22] Basic ingredients of granular computing, i.e., granules, are subsets, classes, and clusters of a universe. They have been considered either explicitly or implicitly in many other fields, such as data and cluster analysis, database and information retrieval, concept formation, and machine learning.[21, 32] In data mining and classification problems, we can view equivalence classes as granules.[27]

The basic guiding principle of fuzzy logic is "*exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness, low solution cost and better rapport with reality.*"[32] GrC offers a more practical philosophy for real world problem solving. In the case of information retrieval on the web, granular computing may provide grouped and personalized view of information retrieval.

The organization of this paper is as follows. In the next section, we will discuss the challenges and opportunities to information retrieval systems in the Internet age. A section that summaries information retrieval support systems is followed. In Section 4, a brief review of granular computing is given. Issues on granulation for web are discussed in Section 5 and 6. Finally, we conclude this article.

2. CHALLENGES AND OPPORTUNITIES TO INFORMATION RETRIEVAL SYSTEMS

Following the study of probability of relevance,[14] information retrieval can be formalized into four models as shown in Figure 1. Let u be a single user, U a class of users, d a document and D a class of documents, these four models and relationships among them can be depicted as in Figure 1. Granulation of a universe involves the decomposition of the universe into parts, or the grouping of individual elements into classes, based on available information and knowledge. Elements in a granule are drawn together by indistinguishability, similarity, proximity or functionality.[32] Using techniques and principles of granular computing, one can study these models from granulation point of view, i.e., grouped and personalized views should be also studied. Model 1 and Model 2 are considered as finer granulation than Model 0. The elements of these two models are granules of the elements of Model 0. The same principle applies to other models. Model 3 is the finest granulation, Model 0 is the coarsest granulation.

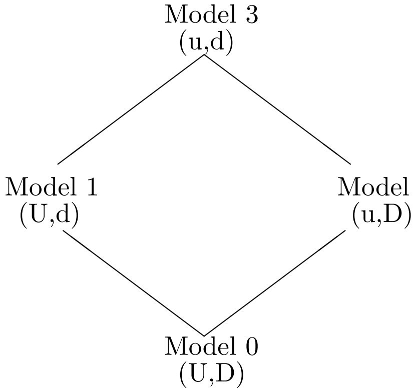

Figure 1. Information retrieval models

Document representation, query formulation, and retrieval functions are fundamental issues in information retrieval study.[19, 20] Based on this classification, we have to design and implement an appropriate scheme to represent the contents of documents, a language to express user queries, and a retrieval function to search for relevant documents. Index terms play the connecting role between documents and user queries. A document is considered to be relevant to a query if the user submitting the query judges the document to be useful.

Different retrieval models have been developed, such as the Boolean, vector space, and probabilistic models[17, 19] as well as recent linguistic and knowledge-based models. The first three models are often referred to as the exact match model; the latter as the best match models.[4] Although they provide us formal and elegant formulations of information retrieval problems, they suffer from several shortcomings. The classical retrieval models are over-simplification of the real world retrieval problem.

Each IR model represents a document simply as a list of terms that appear in the document. The list is often obtained by term weighting schemes. It is typically the results of some statistical analysis of document text. The aim of the weighting is to quantify the degree of breadth or narrowness of the terms. However, the effectiveness of statistical analysis, although providing us useful information, is limited.[19] To avoid this difficulty, one needs to consider the structure information and semantic information of the document. Two levels of structure information, i.e., document level structure and collection level structure, can be used.

The document level structure information shows the connection between components of a document. Such information is now readily available from the online searchable and structured or semi-structured documents prepared using some markup languages. A web page may consist of paragraphes, sentences, words, images, etc. Introduction and conclusion are more likely to appeared in a same scientific article. The collection level structure information shows the connections between documents. The hyper links and search indices are some examples of this structure. More advanced examples are citation and co-citation analysis and subject clustering of documents. papers cite a particular paper can be clustered into same class.

Natural language analysis tools and ontology can also be used to discover both levels of structure information. In many information retrieval systems, all documents are described in the same level of details. The same document representatives are used independent of individual users. The lack of consideration of the diversity, background and intentions of users effects the performance of IR systems. It is expected that multiple and personalized representations of documents may be more effective.

Similar observations can also be made regarding the issue of query formulation. Typically, a query language is used to express user information needs. A query language may be either too restrictive to be very effective, or too complicated to be practically useful. The problem is made worse when a user is not clear what is being searched for. Many search engines accept natural language queries. However, such queries are in many cases translated into a list of keywords or some simple Boolean expression. An effective information retrieval system should support many query languages and tools for expressing user information needs.

Many information retrieval systems use a simple and single retrieval method. In addition, retrieval is based on keyword level matching. Documents containing the keywords appearing in the query are retrieved or ranked higher. Other information that may suggest the relevance of documents is not fully explored. Recent studies on text mining, user behavior analysis, and agent technology may provide potential solutions to such a problem.[25] One needs also to explore the potential of multi-strategy retrieval exemplified by meta-search engines.

In summary, many shortcomings of traditional information retrieval systems and Web search engines make them inadequate to support research using the Web. The advances made in related fields, such as text mining, intelligent agents, and markup languages open new doors to expand information retrieval systems. Information retrieval support systems is a natural evolution from traditional information retrieval systems.

3. INFORMATION RETRIEVAL SUPPORT SYSTEMS

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.[24] We summary IRSS in this section.

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. They new tasks include understanding, analysis, organization, and discovery, in addition to the conventional tasks of search and browsing. IRSS is a natural evolutionary stage from retrieval systems. 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 the literature.[24]

The evolution from IRS to IRSS is due the pitfalls of 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. In addition, an IRS stem from its 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.

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.

We can classify IRSS models into three related types. 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. They allow multi-representation of documents. Granular computing plays an important role in the construction of document models. The retrieval models deals with the search functionality.

The retrieval models provide languages and tools to assist a user to performs 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 deal with the representation and interpretations of results from the search. They 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. 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.

An IRSS focuses on the supporting functionalities of information retrieval. However existing information retrieval systems only focus on the search and browsing functionalities. IRSS are more flexible and combine the functionalities of IRS, Web browser and Web search engines. It is expected that current IRS need to be extended to support more user tasks. 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. The components of an IRSS are very similar to decision support systems and intelligent systems such as data management, model management, knowledge-based management, and user interface subsystems.

4. GRANULAR COMPUTING

There are many fundamental issues in granular computing, such as granulation of the universe, description of granules, relationships between granules, and computing with granules. They may be studied from two related aspects, the construction of granules and computing with granules. The former deals with the formation, representation, and interpretation of granules, while the latter deals with the utilization of granules in problem solving.

The interpretation of granules focuses on the semantic side of granule construction. It addresses the question of why two objects are put into the same granule. It is necessary to study criteria for deciding if two elements should be put into the same granule, based on available information. One must provide necessary semantic interpretations for notions such as indistinguishability, similarity, and proximity. It is also necessary to study granulation structures derivable from various granulations of the universe.[28] The formation and representation of granules deal with algorithmic issues of granule construction. They address the problem of how to put two objects into the same granule. Algorithms need to be developed for constructing granules efficiently.

Computing with granules can be similarly studied from both the semantic and algorithmic perspectives. On the one hand, one needs to interpret various relationships between granules, such as closeness, dependency, and association, and to define and interpret operations on granules. On the other hand, one needs to design techniques and tools for computing with granules, such as approximation, reasoning, and inference.

Let U be a finite and non-empty set called the universe, and let $E \subseteq U \times U$ denote an equivalence relation on U. The pair $apr = (U, E)$ is called an approximation space. The equivalence relation E partitions the set U into disjoint subsets. This partition of the universe is called the quotient set induced by E and is denoted by $U/E = \{[x]_E \mid x \in U\}$, where

$$[x]_E = \{y \mid y \in U, xEy\}, \tag{1}$$

is the equivalence class containing x. The equivalence relation is the available information or knowledge about the objects under consideration. It represents a very special type of similarity between elements of the universe. If two elements $x, y \in U$ belong to the same equivalence class, we say that x and y are indistinguishable, i.e., they are similar. Each equivalence class may be viewed as a granule consisting of indistinguishable elements. It is also referred to as an equivalence granule. The granulation structure induced by an equivalence relation is a partition of the universe. There is a one-to-one correspondence between equivalence relations and partitions of the universe.

A partition is only a very restricted granulation of the universe, in which no overlap between granules is allowed. In general, one can use a covering of the universe to granulate a universe. In this case, a universe is divided into a family of possibly overlap granules. By allowing the overlap between granules, one can put an element into more than one granule. There is no longer a one-to-one correspondence between coverings of a universe and certain type of binary relations on the universe.

The simple one-level granulated views of a universe are based on binary relations representing the simplest type of similarities between elements of the universe. Two elements are either related or unrelated. To avoid such a limitation, multi-level granulation structures can be constructed by putting together simple granulation structures.[23] Each level of the complex structure is a simple granulation structure such as a partition or a covering. A multi-level hierarchical granulation structure can be interpreted and constructed by a nested sequence of binary relations.[11, 23]

5. GRANULATION FOR WEB INFORMATION RETRIEVAL

As discussed above, information retrieval is a term matching process between users and documents. The subprocess on user side is called query formation which forms queries index terms. The subprocess on documents side is called indexing which label documents with index terms too. When terms in a query match documents labelled by the terms, these documents are appeared as the query results. There are four searching spaces, namely, document space, user space, term space and retrieval result space involved in web information retrieval.

5.1. Document Space Granulations

Clustering, classification and association rule analysis are three major data mining techniques. Document clustering is a widely used technique in information retrieval to reduce computational costs and improve retrieval effectiveness. It is a technique for automatically discovering groups of similar documents in a set of documents and grouping the documents by those special topics or clusters. Document clustering may be used as a suggestion for possible relationships in texts or as a hint how to break down search results into smaller pieces. Documents may be clustered in several ways. Content based, query based, and citation based approaches are some examples.

- The most commonly way of clustering is content or topic based approach. Documents with similar content or topic are put into the same cluster.[17, 19] The recent extensive studies and renewed interest on text categorization further explore document clustering.[9] Yahoo! clusters it's collection into Business & Economy, Computers & Internet Internet, News & Media Newspapers, Entertainment, Recreation & Sports, Health, Government, Reginal, etc. based on the content of a web page.

- Another type of document clustering methods is the query oriented document clustering. Documents are clustered based on their joint relevance to a set of queries. That is, documents are put into the same cluster if they tend to be relevant at the same to some queries.

- A similar idea is to use citation and co-citation information for document clustering.[7] Such clustering methods are used in ResearchIndex,[15] in which, for example, co-cited documents are put into a cluster. One may want to find out how a research article is cited in the literature. All the papers that cite the special research article are clustered into a group. Therefore, the extensions of the idea from this article and evolution of the technique can be easily observed by this type of clustering.

- Other methods for clustering documents is based on special characteristics of documents. For example, documents can be clustered based on authors, journals or conference, as in DBLP.[5] Those clustering methods are not only valuable for content oriented retrieval, but also suitable for other special purpose retrieval. In content based document clustering, a collection of documents is divided into clusters such that each cluster consists of similar documents. A center called centroid is constructed for each cluster to represent all the documents in that cluster.[17]

A clustering of documents provide a granulated view of the document collection. One may use either a partition or a covering of the document collection for clustering. Suppose that we obtain a cluster of documents as with the similar contents as animal. A hierarchical clustering of documents is produced by decomposing large clusters into smaller ones. The large clusters offer a rough or abstracted representation of the document. The representation becomes more precise as one moves towards the smaller clusters. A document is described by different representations at various levels. Hence, a cluster-based IR system implicitly employs multi-representation of documents.

Cluster based retrieval is done by comparing a query with the centers of the larger clusters. If the center of the current cluster is sufficiently close to the query, then the query will be compared against the centroid of the smaller clusters at a lower level. In other words, if it is concluded that a document is not likely to be useful using a rough description, then the document will not be further examined using more precise descriptions. Different retrieval methods may also be employed at different levels.

Document clustering only reduces the dimensionality of the document collection while the dimensionality of index terms remains the same. That is, the same number of terms is used for the representation of cluster centers regardless of the level in the document hierarchy. On the other hand, text categorization uses some predefined categories to label or name a cluster, and thus introduces different representations of the same document.

5.2. User Space Granulations

Like the granulation of document space, one can construct granulated views of query user space in several ways, such as content based, document based approaches.[29] Content based query clustering is similar to content based document clustering. The similarity of queries is evaluated based on index terms used by the users or queries. Similar queries are grouped together to represent the needs of a group of users. Content based approaches can be easily extended to cluster users based on user profiles or user logs. On the other hand, document based query clustering methods use the overlap of relevant documents, retrieval results, of queries.[6] Although two queries may not be similar according to their contents, they are still considered to be similar due to a large overlap of relevant documents.

Some recent ideas for query clustering are related to the document based query clustering.[29] They are useful in question answering systems or search engines, such as AskJeeves.[2] Beeferman and Berger[3] suggested to cluster queries by using click-through data, which is implicit relevance information provided by users. Wen et al.[29] combined both content based and document based (through user document clicks) approaches for query clustering. Web usage mining is to extract the usage patterns from Web logs, cookies, etc. in order to use this information in marketing. The first step is to cluster customers by their shopping transactions. It is expected that the customers in the same cluster should have the same shopping preferences.

5.3. Term Space Granulations

The problem of term clustering and its application in information retrieval have been studied by many authors.[16, 17] In a term hierarchy, a cluster may be assigned new terms as labels of the cluster. The new terms are more general than each individual term in the cluster. In general, a multi-level coverings may be more suitable. A more specific term may be described by more than one general terms. For instance, a special term “ice hockey” can be generalized as “winter sports”, “hockey”, “team sports”, “ball games”, etc.

A term hierarchy serves as an effective tool to summarize knowledge about a specific domain. Many domain-specific term hierarchies or concept hierarchies have been used in the organization and retrieval of scientific literature. Examples of term hierarchies are the ACM Classification System[1] and the Mathematics Subject[12] Classification. With such systems, one immediately derives a hierarchical granulation of documents. At each level, documents are described by different terms of different specificities. Documents described by the same terms in the classification system are naturally put into the same cluster.

A main consideration in using term hierarchies is the trade-off relationship between the high dimensionality of index terms and the accuracy of document representation. One may expect a more accurate document representation by using more and specific index terms. However, the increase of the dimensionality of index terms also leads to a higher computational cost.

Term clustering techniques have been used mainly for retrieval, such as query modification, query expansion, and sophisticated retrieval functions.[17] Their use in the granulation of document space and query space has not been fully investigated. The potential of a term (concept) hierarchy, especially its implied knowledge structure, needs to be fully exploited. The fast growing interest on ontology clearly demonstrates a trend in exploring relationship between terms (concepts). Search Engine Watch(seachenginewatch.com) tests search engines size with different type of terms (obscure terms, unusual terms, and popular terms).

Terms space can be granulated by terms appeared in same documents, terms appeared in same query, and terms appeared in a query and a relevant document.

5.4. Retrieval Results Granulations

In many situations, the list of documents returned by information retrieval systems such as search engines is too long and contains duplicate or near-duplicate documents. In order to resolve those problems, many authors suggested and studied the granulation of retrieval results.[10, 18] By clustering retrieval results, one can organize the results and provide coarse-grained summarization to users. Many search engines perform the clustering on a separate machine due to the high volume of searches and queries. The clustering machine receives search engine results as input, creates clusters, and presented them to the user.[33] In fact, many major search engines granulate search results so that you only get one or two top pages per web site. This feature allows us to get more variety and a better chance to find interested pages quickly.

The idea of granulating retrieval results was first studied by Preece.[13, 18] Willet[30] referred to such document clustering as query specific document classification, in contrast to the query independent document granulation discussed earlier. With respect to a particular query, clustering results are more effective.[10, 18]

An important issue in query specific document clustering is to obtain a meaningful description of the derived clusters to be presented to the user. The classical centroid based approach is no longer appropriate. It has been suggested that a few titles and some terms can be used as the description of a cluster.[10] One may also extract some important sentences from the documents in a cluster as a description of the cluster.

When one search for "Regina", for instance, dozens or even hundreds links would come out. The user might search for "University of Regina", "The Regina city", "The Queen", a person's name, or any other cluster related the term "Regina" even the term "reginal". The search engine does not what the user search for. By retrieval results granulation, one may easily pick her interested cluster and browse in lower level details.

6. WEB GRANULATION WITH (SEMI-)STRUCTURED WEB DOCUMENTS

The granulated views of documents and terms can be exploited to discover collection level structures and organize documents accordingly. The document level structure information can be obtained from the use of markup language.

In information retrieval, full text and structured documents have been considered by many authors.[8] With the development of XML (eXtensible Markup Language), there is a growing interest in the organization and retrieval of structured and semi-structured documents. XML is becoming a new standard for data representation and exchange. More XML documents are expected to be available on the Web. A recent findings can be found [26] which describes a GrC model for the organization and retrieval of scientific XML documents.

In XML, the structures and possibly the meaning of data are explicitly indicated by element tags. The structure of a document and element tags are defined through a DTD (Document Type Definitions). For example, a scientific article has a hierarchical (multi-level) granulated description given by the tree structure of DTD. An XML document can be viewed in many different ways by focusing on different tags. For example, one can easily extract only theorems from an XML document. More specifically, an XML document itself can be viewed as the physical view of the documents, and many different logical views can be obtained.

We can explore the rich information provided in XML documents. For example, one can cluster documents using certain tag fields. In this way, different granulated views of the collection can be formed with respect to different users. An XML document collection also allows multi-strategy retrieval. One may use structured queries by focusing on certain tags or perform free text retrieval by simply ignoring all tags. The rich information available in XML documents enable us to extend the traditional functionalities of information retrieval systems.

7. CONCLUDING REMARKS

We discuss the potential applications of data mining techniques for the design of Web based information retrieval support systems (IRSS). In particular, we apply clustering methods for the granulation of different entities involved in IRSS. Two types of granulations, single-level and multi-level granulations, are investigated. We also address issues of granulation for web information retrieval. We address three fundamental considerations in information retrieval, i.e., document representation, query formulation, and retrieval functions techniques, from the Granular Computing point of view. We model web information retrieval searching bases as document space, user query space, term space, and retrieval results space. To introduce granular computing techniques in information retrieval systems, one only needs to examine the problem at a finer granulation level with more detailed information when there is a need or benefit for doing so. The granulated views allow us to focus on the useful structures without looking into too much details. Instead of searching for the optimal solution, one may search for good approximate solutions.

REFERENCES

1. ACM Classification System, http://www.acm.org/class
2. AskJeeves, http://www.askjeeves.com/
3. D. Beeferman and A. Berger, “Agglomerative clustering of a search engine query log”, in *Proceedings of KDD’00*, pp. 407-415, 2000.
4. N. Belkin, and B. Croft, “Information filtering and informatiuon retrieval”, *Communications of the ACM*, 35, pp. 29-37, 1992.
5. DBLP, http://dblp.uni-trier.de
6. L. Fitzpatrick and M. Dent, “Automatic feedback using past queries: social searching”? in *Proceedings of SIGIR’97*, pp. 306-313, 1997.
7. E. Garfield, *Citation Indexing – Its Theory and Application in Science, Technology and Humanities*, New York, John Wiley & Sons, 1979.
8. F.C. Heeman, Granularity in structured documents, *Electronic Publishing* 5, pp. 143-155, 1992.
9. T. Joachims, “Text categorization with support vector machines: learning with many relevant features”, in *Proceedings of the 10th European Conference on Machine Learning*, pp. 137-142, 1998.
10. C. de Loupy, P. Bellot, M. El-Bèze and P.F. Marteau, “Query expansion and classification of retrieved documents”, in *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pp. 382-389, 1998
11. W. Marek and H. Rasiowa, “Gradual approximating sets by means of equivalence relations”, *Bulletin of Polish Academy of Sciences, Mathematics* 35, pp. 233-238, 1987.
12. Mathematics Subject, http://www.ams.org/msc
13. S.E. Preece, “Clustering as output option”, in *Proceedings of the American Society for Information Science*, pp. 189-190, 1973.
14. S.E. Robertson, M.E. Maron and W.S. Cooper, “Probability of relevance: a unification of two competing models for document retrieval”, in *Information Technology: Research and Development* 1, pp. 1-21, 1982.
15. ResearchIndex, http://citeseer.nj.nec.com/cs
16. H. Sakai, K. Ohtake and S. Masuyama, “A retrieval support system by suggesting terms to a user”, in *Proceedings 2001 International Conference on Chinese Language Computing*, pp. 77-80, 2001.
17. G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, New York, McGraw Hill, 1983.
18. A. Tombros, R. Villa and C.J. van Rijsbergen, “The effectiveness of query-specific hierarchic clustering in information retrieval”, *Information Processing and Management* 38, pp. 559-582, 2002.
19. C.J. van Rijsbergen, *Information Retrieval*, London, Butterworths, 1979.
20. S.K.M. Wong and Y.Y. Yao, “On modeling information retrieval with probabilistic inference”, *ACM Transactions on Information Systems* 13, pp. 38-68, 1995.
21. Y.Y. Yao, “Relational interpretations of neighborhood operators and rough set approximation operators”, *Information Sciences* 111, pp. 239-259, 1998.
22. Y.Y. Yao, “Granular computing: basic issues and possible solutions”, in *Proceedings of the 5th Joint Conference on Information Sciences* Vol. I, pp. 186-189, 2001.

23. Y.Y. Yao, “Information granulation and rough set approximation”, *International Journal of Intelligent Systems* 16, pp. 87-104, 2001.
24. Y.Y. Yao, Informaiton retrieval support systems, *Proceedings of the 2002 IEEE World Congress on Computational Intelligence* (2002) pp. 773-778.
25. Y.Y. Yao, H.J. Hamilton and X. Wang, “PagePrompter: an intelligent Web agent created using data mining techniques”, in *Proceedings of International Conference on Rough Sets and Current Trends in Computing*, LNAI 2475, pp. 506-513, 2002.
26. Y.Y. Yao, K. Song and L.V. Saxton, “Granular computing for the organization and retrieval of scientific XML documents”, *Proceedings of the Sixth International Conference on Computer Science and Informatics*, pp. 377-381, 2002.
27. Y.Y. Yao, and J.T. Yao, “Granular computing as a basis for consistent classification problems”, in *Proceedings of PAKDD’02 Workshop on Toward the Foundation of Data Mining*, pp. 101-106, 2002.
28. Y.Y. Yao and N. Zhong, “Granular computing using information tables”, in T.Y. Lin, Y.Y. Yao and L.A. Zadeh, (eds.) *Data Mining, Rough Sets and Granular Computing*, Heidelberg, Physica-Verlag, pp. 102-124, 2002.
29. J.R. Wen, J.Y. Nie and H.J. Zhang, “Query clustering using user logs”, *ACM Transactions on Information Systems* 20, pp. 59-81, 2002.
30. I. Willett, “Query specific automatic document classification”, *International Forum on Information and Documentation* 10, pp. 28-32, 1985.
31. I. Willett, “Recent trends in hierarchic document clustering: a critical review”, *Information Processing and Management* 24, pp. 577-597, 1988.
32. L.A. Zadeh, “Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic”, *Fuzzy Sets and Systems* 19, pp. 111-127, 1997.
33. O. Zamir and O. Etzioni, “Web document clustering: A feasibility demonstration”, in *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’98)*, pp. 46-53, 1998.