2008

Information Granulation for the Design of Granular Information Retrieval Systems

Raymond Y.K. Lau
City University of Hong Kong, raylau@cityu.edu.hk

Chapmann C.L. Lai
City University of Hong Kong, chunllai@cityu.edu.hk

INFORMATION GRANULATION FOR THE DESIGN OF GRANULAR INFORMATION RETRIEVAL SYSTEMS

Granulation de l'information pour la conception d'un système granulaire d'extraction d'information.

Completed Research Paper

Raymond Y.K. Lau
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon,
Hong Kong SAR
raylau@cityu.edu.hk

Chapmann C.L. Lai
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon,
Hong Kong SAR
chunllai@cityu.edu.hk

Abstract

With the explosive growth of the amount of information stored on computer networks such as the Internet, it is increasingly more difficult for information seekers to retrieve relevant information. Traditional document ranking functions employed by Internet search engines can be enhanced to improve the effectiveness of information retrieval (IR). This paper illustrates the design and development of a granular IR system to facilitate domain specific search. In particular, a novel computational model is designed to rank documents according the searchers' specific granularity requirements. The initial experiments confirm that our granular IR system outperforms a classical vector-based IR system. In addition, user-based evaluations also demonstrate that our granular IR system is effective when compared with a well-known Internet search engine. Our research work opens the door to the design and development of the next generation of Internet search engines to alleviate the problem of information overload.

Keywords: Information retrieval, information granulation, granular computing, Internet search engines, information overload

Résumé

Ce papier présente la conception et le développement d'un nouveau système granulaire d'extraction d'information afin de faciliter les recherches dans un domaine donné. En particulier, un nouveau modèle calculatoire est conçu pour classer les documents selon la granularité spécifique des requêtes. Les expérimentations initiales confirment que notre système granulaire d'extraction d'informations est efficace, en comparaison avec un moteur de recherche bien connu d'internet.

Abstract in Native language

□□
□□
□□□□□□□□□□□□□□□□□□□□

Introduction

Because of the rapid growth of the amount of information available on electronic networks such as the Internet, it is becoming increasingly more difficult for information seekers to locate relevant information. This leads to the so-called problem of information overload (Lau et. al. 2008; Niu and Winter 2006; Yao 2002). Classical similarity-based IR techniques (Salton et. al. 1975; Salton 1990) have been applied to information search on the World Wide Web (Web). In addition, popularity-based IR method such as the PageRank algorithm has been developed to augment similarity-based IR on the Web (Page et. al. 1998). One implicit assumption behind the document ranking functions of existing Internet search engines is that popularity is closely correlated with relevance. Unfortunately, the correlation between popularity and relevance is very weak for newly created pages that do not have many inlinks (Mowshowitz and Kawaguchi 2002). The fact is that while the recall of information is facilitated by powerful Internet search engines, the precision of the search results is still relatively low (Lawrence 2000).

Recent advance in “granular computing” (Bargiela and Pedrycz 2008; Yao 2005; Yao 2002) sheds light on developing more effective IR systems to alleviate the problem of information overload. The granular computing paradigm emphasizes the effective use of levels of “granularity” or abstraction to systematically analyze, represent, and solve real-world problems (Bargiela and Pedrycz 2008; Yao 2005). In granular computing, information granulation refers to the computational processes of generating and presenting levels of granularity of information to facilitate problem solving (Yao-JT 2005; Zadeh 1979). In the context of IR, we apply the concept of information granulation to design a granular IR system which can estimate the granularity of documents (e.g., general vs. specific documents) and rank these documents with respect to information seekers’ specific granularity requirements. By applying the design science research methodology (Hevner et al. 2004), we will illustrate the design, development, and evaluation of the proposed granular IR system in this paper.

The Needs for Granular IR Systems

Existing Web-based IR systems such as Google Maps[1] supports information granulation for a special kind of information objects (i.e., geographical maps). As shown in Figure 1, a slider bar (the granularity control) is provided to the searchers so that they can view a geographical location at different levels of granularity. For instance, the left hand side of Figure 1 shows the city centre of Paris at a high abstraction level, whereas the right hand side depicts the same place with finer details. Unfortunately, information granulation is not supported for general Web document retrieval. Imagine that it would be nice to have a granularity control bar beside the Google query box to facilitate the retrieval of Web documents at different levels of granularity. The granularity of a document refers to the levels of details (i.e., the specificity) of information contained in the document. Since specificity is the antonym of generality, we will measure document granularity (an attribute) in terms of document specificity (attribute value) throughout this paper.

a. A High Level View of Paris | **b. A Finer View of Paris**

Figure 1. An Example of Information Granulation

[1] http://maps.google.com/

In typical IR situations, information seekers often need to retrieve information with different levels of details to fulfill their specific information needs. For instance, a student may need to develop an overview of a research area by reading general reviews or introductory documents (e.g., an introduction to information systems). In another case, a professor may want to read articles with sufficient details for a specific topic (e.g., design science for information systems) for launching a new research project. Because of the sheer volume of documents archived on the Web and digital libraries, it is not practical to manually label "general" or "specific" documents individually. In fact, it is almost impossible to assign a static granularity label to a document because granularity is not entirely determined by the content of the document. To certain extent, the granularity of a document depends on the knowledge states and the tasks at hand of a searcher. For example, the document about "an introduction to information systems" may still be considered containing too much technical details for a high school student. Therefore, there is a pressing need for the development of granular IR systems which can automatically retrieve documents with respect to the information seekers' specific granularity requirements.

A Holistic Document Ranking Function

Document ranking is the fundamental function of any IR systems such as the Internet search engines (Haveliwala 2003). A document ranking function assigns a score to a document according to some objective criteria such as the similarity between the document and the query. The cosine similarity function is one of the widely used document ranking functions (Salton et. al. 1975). An IR system usually ranks and lists documents (e.g., Web pages) in descending order of similarity scores (Salton 1990). With the invent of the PageRank algorithm and its variants (Haveliwala 2003; Page et. al. 1998), most Internet search engines employ a hybrid similarity-based and popularity-based mechanism to rank Web documents these days. An effective ranking function is essential for the success of Internet search engines since information seekers rarely review Web documents beyond the first page of a result set (Granka et. al. 2004). To improve the effectiveness of IR in general and Web search in particular, we propose a novel document ranking function which adopts a holistic view of document ranking. As depicted in Figure 2, we argue that an effective document ranking function should take into account three orthogonal dimensions, namely similarity, popularity, and granularity.

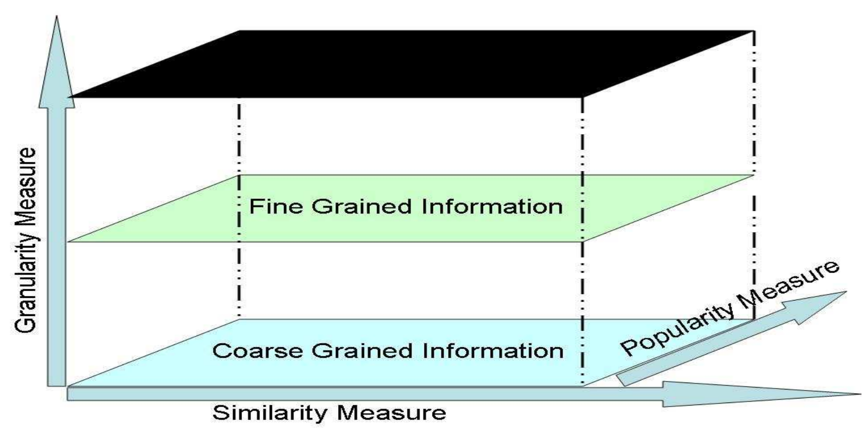

Figure 2. Three Dimensions of a Holistic Document Ranking Function

Main Contributions

The theoretical contribution of our research work is that the granular computing paradigm is exploited in the context of IR. Based on the concept of information granulation, a granular IR system is designed and developed. In particular, a novel computational model for automatically estimating document (query) granularity is developed to enhance domain specific search. From the practical stand point, our research work opens the door to the development of the next generation of Internet search engines to alleviate the problem of information overload.

Outline of the Paper

The rest of the paper is organized as follows. The next section highlights the architectural design of a granular IR system, and then followed by a discussion of related research work. A computational model for implementing the holistic document ranking function is then illustrated. The quantitative evaluations of the granular IR system based on system-oriented and user-oriented experiments are reported afterwards. Finally, we offer concluding remarks and describe future direction of our research work.

The Architectural Design of a Granular IR System

The general system architecture of our granular IR system[2] is depicted in Figure 3. An information seeker first translates their implicit information needs (including granularity requirement) into explicit queries. Recurring queries are often stored in a user profile within the granular IR system. On the other hand, information objects (e.g., Web pages) from specific information sources such as the Internet are characterized by a particular indexing scheme. These document characterizations are also stored in the local cache of the granular IR system. The document ranking mechanism of the granular IR system computes an aggregated document score for each document by taking into account three aspects, namely similarity, popularity, and granularity with respect to the given query. The granular presentation layer of the IR system will generate the appropriate presentation formats (e.g., ranked list of documents or clusters of documents) with respect to the specific preferences of individual user or group of users. As a result, information seekers can retrieve information with the levels of details and formats they prefer. After reviewing the information objects, information seekers may provide relevance feedback about the content and the presentation format of the delivered documents to the learning and adaptation mechanism. Thereby, both the content and the presentation format of documents can be improved in subsequent round of information search. Our prototype system was developed using Java (J2SE v 1.4.2), Java Server Pages (JSP) 2.1, and Servlet 2.5.

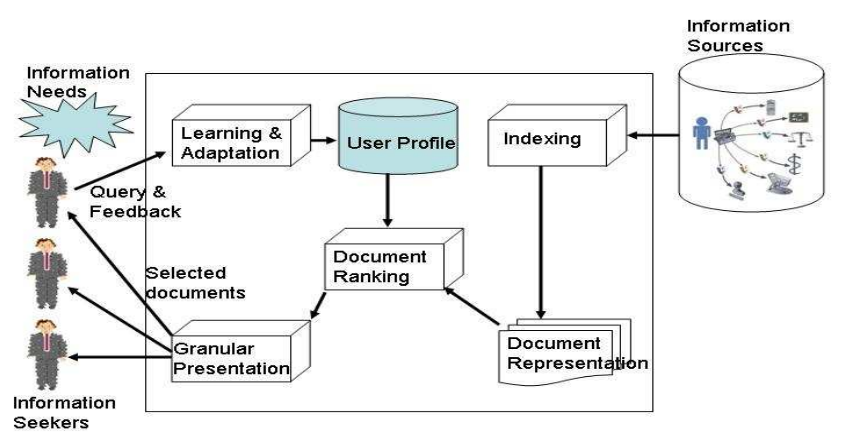

Figure 3. The General System Architecture of a Granular IR System

In this paper, we will only focus on the document ranking mechanism of the granular IR system. In order to estimate whether a document contains specific or general information, the granular IR system needs to consult a domain ontology which captures the knowledge about a specific application area. Ontology is generally considered as a formal specification of conceptualization which usually takes the form of taxonomies of concepts (Gruber 1993). One well-known example is the Medical Subject Headings (MeSH)[3] domain ontology which captures rich taxonomical knowledge about the life-science domain.

[2] http://144.214.55.127:8080/ong/tiger/OntologyUI.jsp

[3] http://www.nlm.nih.gov/mesh/

Related Research

As granular computing is a relatively new system development methodology, there are few studies to examine the design and development of granular IR systems. Yao (2002) is probably the first researcher to explore the idea of granular computing in the context of IR. It was proposed that an IR support system should exploit document space granulations (e.g., document clustering), user space granulations (e.g., grouping similar queries into a group user profile), term space granulations (e.g., grouping terms by specificity or generality), and retrieval result granulations (e.g., clustering the result sets) to develop effective IR systems for an individual or group of information seekers (Yao 2002). However, the idea of applying the granular computing methodology to IR remains as a conceptual discussion rather than a concrete system design and development work. Our research extends the idea of granular information retrieval support system by designing, implementing, and evaluating a prototype granular IR system. In particular, term space granulation is exploited to construct a computational model to estimate the granularity of documents and queries.

Documents and result sets clustering (i.e., retrieval result granulation) have been examined by many researchers in the field of IR (Roussinov and Chen 2001; Buyukkokten et. al. 2002). For instance, different textual units of Web documents were identified, grouped, and summarized to produce satisfactory displays on small handheld devices (Buyukkokten et. al. 2002). Thereby, users could quickly absorb the most important information of a Web page even though the physical display size of a handheld device was quite limited. Our granular IR system not only groups documents into clusters but it can also assess the granularity of each individual document.

Roussinov and Chen (2001) employed the Kohonen's self-organizing maps (SOM) to cluster search results returned from the Alta Vista search engine. Through such a clustered result set (i.e., a high level abstraction view), information seekers found it easier to understand the semantic contents of the returned Web pages. Our granular IR system also makes use of existing application programming interfaces (APIs) provided by an Internet search engine. However, our granular IR system supports both term space granulation and result set granulation. In addition, to fulfill an information seeker's specific granularity requirement, a more fine grained query and document granularity computation is supported.

It was argued that the readability of a document could be assessed quantitatively and objectively (Yan et. al. 2006). Accordingly, an ontology-based computational model was developed to assess the readability of documents. The assumption was that if a document contained some terms which appeared in a concept hierarchy (i.e., ontology) of a specific domain, the readability score of that document decreased. They believed that it was generally more difficult for the readers to digest domain specific terminologies (Yan et. al. 2006). Moreover, the deeper the terms were found in the concept hierarchy, the less readable the document would be. Our granular IR system also makes use of domain ontology to objectively assess the granularity (e.g., the specificity) of documents. However, we avoid modeling the more subjective issue of readability which may not be quantified based on domain ontology alone.

Zhou et. al. (2006) examined the issues of information specificity and information generality in the context of ontological user profiling and users' search intension modeling. Their computational model for estimating information specificity and information generality was based on Dempster-Shafer (D-S) theory of evidence. In particular, the specificity measure was developed based on the belief function of the D-S theory, whereas the generality measure was constructed based on the plausibility function of the D-S theory. Instead of employing the D-S theory of evidence to estimate document granularity, we develop an efficient ontological computational model to estimate document granularity.

It was proposed that complementing a search query with document genre requirement might improve the performance of search engines (Ferizis and Bailey 2006). Document genre refers to the types of documents, such as a diary or a professional report. Either comprehensive linguistic analysis or term frequency based techniques could be applied to estimate document genre (Ferizis and Bailey 2006). Our work focuses on document granularity which can be taken as a special kind of document genre e.g., general vs. specific documents. We believe that efficient computation of document granularity, a simpler form of document genre, could be achieved under an online interactive environment. Therefore, our method is readily applied to enhance existing Internet search engines.

A Computational Model for Granular IR

In this section, we first explain the intuitions behind our granular IR model. Then, the computational model for estimating document and query granularity is illustrated. Finally, the holistic document ranking function is defined formally.

The Notion of Document Granularity

Intuitively, a document is considered specific if it contains some domain specific terminologies. For instance, comparing a document about "diseases" and another document about "pneumonia", the latter is probably considered to be more specific than the former because it refers to a specific kind of disease using formal terminology. In addition, a document with specific terminologies (e.g., "conjunctivitis") is considered to contain more details than another document with general terms (e.g.., "eye infection"). With reference to Figure 4 which shows a segment of the MeSH domain ontology, we know that "conjunctivitis" is one specific kind of "eye infection", and so "conjunctivitis" is a more specific terminology than "eye infection" is. We propose the notion of "terminological specificity" to measure the proportion of domain specific terminologies appearing in a document.

The previous examples show that it is possible to develop a computational model for estimating document granularity with the help of domain ontology. For instance, the MeSH ontology provides a controlled vocabulary of 21,836 terms for the life science domain. The MeSH domain ontology has been applied to index PubMed[4], a popular online database of 11 million medical and health related citations and abstracts. Throughout this paper, the MeSH domain ontology will be used to illustrate the main ideas of our granular IR system, and it will also be used to evaluate the effectiveness of our prototype system.

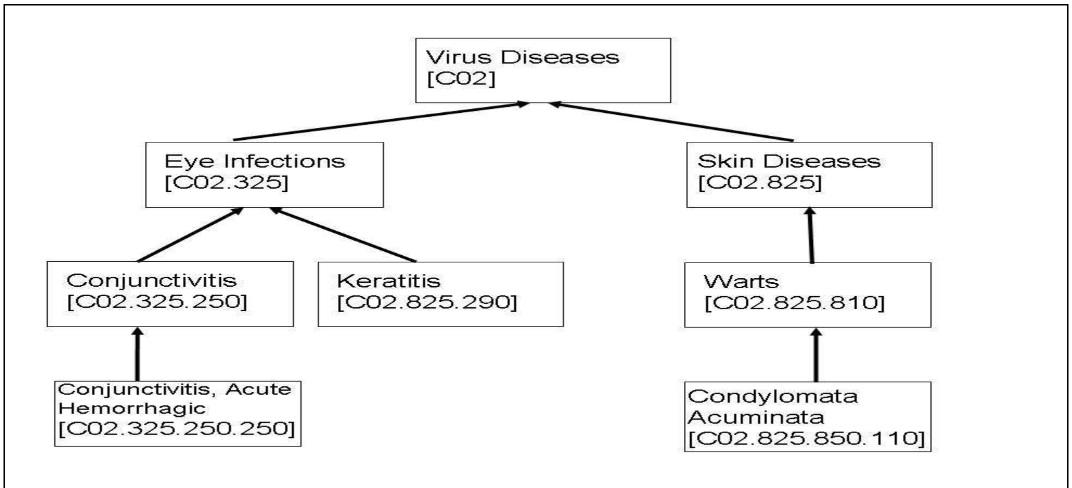

Figure 4. A Segment of the MeSH Domain Ontology

Apart from terminological specificity, there is another factor which also contributes to document specificity. Consider that two documents containing terminologies of the same level as encoded in a concept hierarchy may still demonstrate different specificity. With reference to Figure 4, a document about "conjunctivitis" and "keratitis" is probably considered more specific than another document about "conjunctivitis" and "warts". The reason is that both "conjunctivitis" and "keratitis" are specifically referring to "eye infection", whereas "warts" is about skin disease. This gives rise to the notion of "referential specificity" which refers to the cohesion of the terminologies covered by a document. In this paper, we argue that the granularity of a document can be estimated based on two orthogonal dimensions, that is, terminological specificity and referential specificity.

Terminological specificity and referential specificity can be estimated with reference to a conceptual hierarchy which encodes the specialization relationships from a parent root node to the children nodes. Conceptual hierarchies (ontologies) have been manually constructed for general information domains or for a specific subject domain. An example of the former kind of ontology is the Library of Congress Subject Headings (LCSH)[5], and an example for

---

[4] http://www.ncbi.nlm.nih.gov/pubmed/

[5] http://classificationweb.net

the latter is the Medical Subject Headings or the ACM computing classification system[6]. It should be noted that machine readable format of a portion of LCSH is made available through the World Wide Web Consortium's (W3C) Simple Knowledge Organization System (SKOS) project at (http://lcsh.info/sh85118553). This sheds light on the technical feasibility of applying the proposed granular IR system to search for documents of various domains. In addition, recent advance in automated methods for domain ontology construction (Griffiths et. al. 2007; Lau et. al. 2007) also extends the applicability of the proposed granular IR systems to theoretically unlimited information domains.

Document Representation

Each document was pre-processed according to traditional IR techniques (Salton et. al. 1975; Salton 1990). For example, each document was parsed according to a variant of the SMART stop word file (Salton 1990) to remove semantically insignificant terms such as articles, prepositions, and others. Then, the Porter stemming algorithm (Porter 1980) was applied to compute the root form of terms. For example, the words "computer", "computers", and "computing" were all converted to the stem "comput" after applying the stemming procedure. A Term Frequency Inverse Document Frequency (TFIDF) like heuristic was applied to identify the most representative terms of the document (Salton et. al. 1975). The top *N* terms with the highest TFIDF weights were then selected to represent the semantic content of the corresponding document. Similarly, a query was treated as a short document and the same document pre-processing method was applied to the query. For any domain ontology utilized by our Granular IR system, the same stop word removal and stemming procedures were applied to the vocabulary of the ontology. A standard substring function could then be applied to check if a document containing the terminologies of the ontology.

The Similarity Ranking Function

The relevance of a document d to a query q can be estimated based on a similarity function (Salton 1990). In the vector space model (Salton et. al. 1975; Salton 1990), both documents and queries are represented by the corresponding vectors of term weights. A document is then ranked according to the similarity between the document vector $\vec{d}$ and the query vector $\vec{q}$. A widely used similarity function is the cosine function (Salton 1991) and it is defined by:

$$Sim(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{n} w(t_i^q) \times \sum_{i=1}^{n} w(t_i^d)}{\sqrt{\sum_{i=1}^{n}\left(w(t_i^q)\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(w(t_i^d)\right)^2}} \tag{1}$$

where $w(t_i^q)$ and $w(t_i^d)$ are the weight (e.g., TFIDF weight) of the *i* term in the query vector $\vec{q}$ and the document vector $\vec{d}$ respectively. The term *n* represents the dimension (e.g., the number of term weights) of the vectors. The range of the cosine similarity function $Sim(\vec{q},\vec{d})$ falls in the unit interval [0,1].

Terminological Specificity

Terminological specificity of a document can be estimated according to the coverage and the level of specificity of the terms contained in the document. With reference to a domain ontology such as the one depicted in Figure 4, the more low-level terminologies appear in a document, the more specific the document will be. The function $TS(d)$ is proposed to measure the terminological specificity of a document by computing the average depths of all the terms appearing in the document. The depth of a terminology is measured in terms of the distance between the terminology node and the root node. The depth of a term not appearing in the ontology is assumed zero (i.e., it has no specific meaning according to the domain ontology). Terminological specificity $TS(d)$ is formally defined by:

---

[6] http://www.acm.org/class/1998/

$$TS(d) = \frac{\sum_{t \in MC} depth(t)}{max(ran(depth)) \times |MC|} \qquad (2)$$

where MC refers to the set of matching terminologies found in a document d, and $|MC|$ is the cardinality of the set MC. The function $depth(t)$ returns the depth of the terminology t with respect to a concept hierarchy. The operator ran returns the range of a function. The normalization factor $\frac{1}{max(ran(depth))}$ is applied to the $TS(d)$ function such that its range falls in the unit interval.

Referential Specificity

Referential specificity is another major factor which determines the granularity of a document. If the terms of a document are more cohesive (e.g., they refer to the same topic), the document is considered more referentially specific. The concept of referential specificity also applies to documents which cover several information topics (domains). Consider that a universal concept hierarchy such as LCSH is adopted as the source of reference, the distance among the terminologies of arbitrary domains can be estimated through traversing different sub-trees of the universal concept hierarchy. The referential specificity of a document is inversely proportional to the distances among its constituent terms as encoded in a concept hierarchy. More specifically, the referential specificity of a document d is defined by:

$$RS(d) = \frac{\chi}{\sum_{t_i \in MC, t_j \in MC, t_i \neq t_j} dist(t_i, t_j)} \qquad (3)$$

$$\chi = \frac{|MC| \times (|MC| - 1)}{2} \qquad (4)$$

where *MC* is the set of matching terminologies found in a document. The function $dist(t_i, t_j)$ returns the distance between two terminologies t_i and t_j with reference to a concept hierarchy. For instance, if t_i is directly linked to t_j in a concept hierarchy, their distance is 1. The term $\chi = \frac{|MC| \times (|MC| - 1)}{2}$ returns the maximum number of terminology pairs (t_i, t_j) constructed from the set *MC*.

With reference to Figure 4, consider a document d_1 = {warts, condylomata acuminata} and another document d_2 = {warts, conjunctivitis}. In this example, the value of χ is $\frac{2 \times 1}{2} = 1$ for both documents. For document d_1, the distance between "warts" and "condylomata acuminate" is 1. Nevertheless, for document d_2, the distance between "warts" and "conjunctivitis" is 4 (i.e., 4 links away). Accordingly, the referential specificity of document d_1 is $\frac{1}{1} = 1$, and the referential specificity of document d_2 is $\frac{1}{4} = 0.25$. The referential specificity of document d_1 is higher than that of d_2 because the terminologies appearing in d_1 are more cohesive. In fact, document d_1 is specifically about skin diseases, whereas document d_2 covers both skin disease and eye infection. As can be seen, the range of the function $RS(d)$ falls in the unit interval. Our current method does not consider the directions of the links encoded in the conceptual hierarchy. Whether taking into account the directions of the semantic links can further improve the effectiveness of our current method will be a subject for our future research.

Document and Query Specificity

By taking into account both terminological specificity and referential specificity, the specificity of a document can be estimated according to Equation 5. If we treat a query q as a short document and apply the same approach to estimate its specificity, the query specificity of q can be derived from Equation 6. The weight factor $\varpi_d \in [0,1]$ controls the relative importance of terminological specificity and referential specificity in estimating the overall document specificity. Similarly, the weight factor $\varpi_q \in [0,1]$ is used to tune the query specificity measure. The range of document specificity or query specificity falls into the unit interval. The automated means of computing query specificity (Eq. 6) is one of the ways to deal with the variance of the perceived granularity of documents among different information seekers. For instance, while an information seeker perceives one document as specific, another person may think that the same document is relatively general. Nevertheless, such a variance will be reflected by the different usage of terminologies in the respective queries. As a result, the variance of information seekers' perceived document granularity due to different knowledge states or tasks at hand can be captured by our system. Our granular IR system also allows information seekers to explicitly specify their granularity requirements by using a granularity control bar (like the Google Maps slider bar).

$$DS(d) = \varpi_d \times TS(d) + (1 - \varpi_d) \times RS(d) \tag{5}$$

$$QS(q) = \varpi_q \times TS(q) + (1 - \varpi_q) \times RS(q) \tag{6}$$

Re-Ranking Documents by Granularity

To implement a holistic ranking function, our granular IR system can re-rank documents after applying a similarity-based ranking function like the one defined in Equation 1. Our re-ranking method is also applicable to Internet search engines which employ both similarity measure and popularity measure such as the PageRank algorithm (Page et. al. 1998) to rank Web documents. Our novel ranking function as defined in Equation 7 takes into account the "granularity gap" between a given query and an arbitrary document. The basic intuition is that if there is a large granularity gap between a query and a document (e.g., a specific query versus a general document), the initial similarity or popularity score of the document should be adjusted (e.g., lowered). The reason is that the document is unlikely to meet the information seeker's granularity requirement. On the other hand, if the granularity gap between a query and a document is small, little or no adjustment of the similarity or the popularity score is required. The granularity gap between a query q and a document d is estimated based on the absolute difference between $DS(d)$ and $QS(q)$. The parameter ϖ_G controls the relative weight of the granularity factor when documents are ranked.

$$GScore(d,q) = Sim(d,q) - \varpi_G \times \left| DS(d) - QS(q) \right| \tag{7}$$

Experiments and Results

A two-stage evaluation procedure was applied to verify the effectiveness of our granular IR system. The first stage was a system-oriented benchmark test; we compared the effectiveness of the aggregated similarity and granularity document ranking method with the classical cosine similarity ranking method based on the TREC-AP collection (Hull 1998) and the OHSUMED collection (Hersh et. al. 1994). At the second stage, we conducted user-based empirical studies to evaluate the perceived relevance of the highly ranked documents returned by the granular IR system.

Benchmark Tests

Our experimental procedure was based on the routing task employed in the TREC forum (Hull 1998). Essentially, a set of pre-defined topics (i.e., queries) was selected to represent the hypothetical user information needs. By invoking the respective IR systems (e.g., the granular IR system and the baseline system), documents from the benchmark corpora were ranked according to their relevance to the queries. Standard performance evaluation measures such as precision, recall, mean average precision (MAP) were applied to assess the effectiveness of the respective IR systems (Hersh et. al. 1994). Precision is the fraction of the number of retrieved relevant documents to the number of retrieved documents, whereas recall is the fraction of the number of retrieved relevant documents to the number of relevant documents (Van Rijsbergen 1979). In particular, we employed the TREC evaluation package available at Cornell University[7] to compute all the performance data. The TREC-AP collection comprises the Associated Press (AP) newswires covering the period from 1988 to 1990 with a total number of 242,918 documents (Hull 1998), whereas the OHSUMED corpus (Hersh et. al. 1994) is a set of 348,566 documents consisting of titles and abstracts from 270 medical journals over a five-year period from 1987 to 1991. These benchmark collections also contain the relevance judgment files that define which documents are relevant or non-relevant for a particular information topic.

Our first benchmark test involved the OHSUMED collection which consisted of 106 topics (i.e., queries). Each topic contained both the patient's information and the physician's information. As there was no relevant document for topics 8, 28, 49, 86, and 93, these topics were excluded from our experiment. A baseline system was developed based on the classical vector space model (Salton et. al. 1975; Salton 1990); the similarity scores of the OHSUMED documents were computed according to Equation 1. With respect to each test query, the first 1,000 documents from the ranked list were used to evaluate the performance of an IR system. Our granular IR system employed both the aggregated document ranking function (Equation 7) to rank documents. The query specificity of each OHSUMED query was computed according to Equation 6. For all the experiments reported in this paper, the parameters $\varpi_d = \varpi_q = 0.41$ and $\varpi_G = 0.83$ were used. These system parameters were estimated based on the pilot tests which involved a subset of the OHSUMED test queries.

Table 1. Results of the OHSUMED Benchmark Test

Recall Level	Baseline System		Granular IR System		*t*-statistics df(100)	*p* values
	Mean	STD	Mean	STD		
0	0.5469	0.3595	0.6011	0.3057	2.205	<.01**
0.1	0.3571	0.3039	0.381	0.2912	2.194	<.01**
0.2	0.2739	0.2416	0.2922	0.2651	1.759	<.05*
0.3	0.2044	0.2775	0.2185	0.2534	1.534	=.06
0.4	0.1463	0.1804	0.1642	0.1922	1.758	<.05*
0.5	0.1179	0.1839	0.1381	0.1931	2.089	=.01**
0.6	0.0896	0.1315	0.1064	0.1594	1.603	=.05*
0.7	0.0623	0.1373	0.0721	0.1359	1.245	=.11
0.8	0.0444	0.1114	0.0502	0.0906	0.815	=.21
0.9	0.0223	0.0355	0.0267	0.0392	0.677	=.26
1	0.0024	0.0093	0.0038	0.0098	0.116	=.45
Overall MAP	0.1734	0.1825	0.1868	0.1811	1.527	=.06
Overall Δ%			**7.70%**			

Table 1 shows the means and standard deviations of the precision values achieved by the baseline system and the granular IR system at every recall level. A mean precision value was computed based on the interpolated precision values achieved by an IR system from all the test topics. This is a standard evaluation procedure for the TREC routing tasks (Hull 1998). On the other hand, the overall MAP (as shown at the bottom of Table 1) was computed based on the non-interpolated precision value achieved by an IR system for each topic. At every recall level, we tried to test the null hypothesis (H_{null}: $\mu_{Granular} - \mu_{Baseline} = 0$) and the alternative hypothesis ($H_{alternative}$: $\mu_{Granular} - \mu_{Baseline} > 0$), whereas $\mu_{Granular}$ and $\mu_{Baseline}$ represented the mean precision values achieved by the granular IR

[7] ftp://ftp.cs.cornell.edu/pub/smart/

system and the baseline IR system respectively. The last two columns of Table 1 show the results of our paired one tail *t*-test. An entry in the last column marked with (**) indicates that the corresponding null hypothesis is rejected at the 0.01 level of significance or below, whereas an entry marked with (*) indicates that the null hypothesis is rejected at the 0.05 level of significance or below. As shown in Table 1, the granular IR system achieves better precision at all levels of recall, and there are statistically significant improvement at several recall levels. Figure 5 depicts the precision-recall graphs for the two systems. It is shown that the granular IR system outperforms the baseline system at all levels of recall. In terms of MAP, the granular IR system achieves a 7.7% overall improvement (Δ%); such an improvement is generally considered a good achievement in the field of IR (Ponte and Croft 1998).

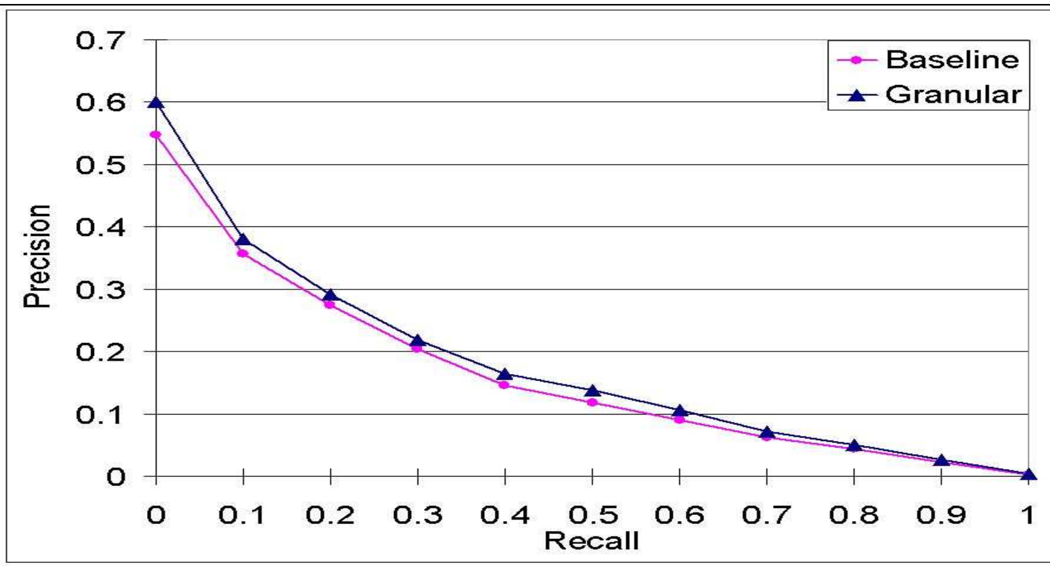

Figure 5. The Precision-Recall Graphs Based on the OHSUMED Benchmark Corpus

Table 2. Results of the TREC-AP Benchmark Test

Recall Level	Baseline System		Granular IR System		*t*-statistics df(9)	*p* values
	Mean	STD	Mean	STD		
0	0.5223	0.1001	0.5942	0.1011	5.279	<.01**
0.1	0.3035	0.1730	0.3762	0.1849	4.905	<.01**
0.2	0.2325	0.1888	0.2843	0.2020	4.584	<.01**
0.3	0.1994	0.1706	0.2491	0.2183	3.723	<.01**
0.4	0.1687	0.1528	0.2255	0.2134	3.340	<.01**
0.5	0.1240	0.1115	0.2010	0.2021	2.732	=.01**
0.6	0.0869	0.0791	0.1527	0.1375	2.557	<.05*
0.7	0.0644	0.0743	0.1058	0.1088	2.061	<.05*
0.8	0.0390	0.0759	0.0603	0.0568	1.086	=.15
0.9	0.0103	0.0235	0.0224	0.0313	1.180	=.13
1	0.0047	0.0026	0.0159	0.0252	1.537	=.08
Overall MAP	0.1519	0.1133	0.1735	0.1269	1.813	=.05*
Overall Δ%			**14.22%**			

To evaluate the effectiveness of our granular IR system for non-medical domains, we adopted the TREC-AP collection for our second benchmark test. We randomly selected ten TREC topics such as "Antitrust", "Acquisitions", "AIDS treatments", "Space Program", "Water Pollution", "Japanese Stock Market Trends", "New Medical Technology", "Influential Players in Multimedia", "Impact of Religious Right on U.S. Law", and "Computer Virus Outbreaks" for our experiment. Each of these topics contains relevant documents. A query was constructed based on the title and the narrative field of the topic. For each TREC topic, we employed our fuzzy ontology extraction algorithm (Lau et. al. 2007) to automatically generate a domain ontology based on the full-text description of the topic. The average depth of the automatically generated ontology was 4.1 and the number of nodes captured in an ontology was 22.9 on average. Similar to the OHSUMED benchmark test, the TREC routing tasks were conducted by the granular IR system and the baseline system respectively. The performance data as computed by the TREC evaluation package is tabulated in Table 2. The granular IR system achieves better precision at all levels of recall, and there are statistically significant improvement at most levels. In terms of MAP, the granular IR system achieves a 14.22% overall improvement (Δ%), and such an improvement is shown to be statistically significant.

In order to carry out a deeper analysis of the experimental results, we manually verified the ranked lists of documents generated by the granular IR system and the baseline system respectively. We use TREC topic 18 "Japanese Stock Market Trends" of the TREC-AP collection as an example to explain why our granular IR system can improve IR effectiveness. The titles of the topic and the extracts of the corresponding documents are listed below:

TREC Topic 18	Title: ***Japanese Stock Market Trends*** Narrative: To be relevant, a document will identify ***trends*** in the ***Japanese*** (***Nikkei***, ***Tokyo***) ***stock market***
AP890524-0295	Heading: "A Cheery Message From Wall Street?" Details: "If you believe the optimists of Wall Street, the ***stock market*** is sending out a happy message for the American economy these days. …………… The average, which is the oldest and best known measure of ***market trends*** ………….."
AP890109-0325	Heading: "***Stocks*** Leap In ***Tokyo*** After Hirohito's Death" Details: "***Share*** prices posted their biggest gain in a year and the ***Nikkei Stock*** Average set a record on Monday, the first day of trading since Emperor Hirohito's death of cancer on Saturday……………"

The highlighted terms appearing in the topic and the documents are captured in the automatically generated ontology. As both the title field and the narrative field of a TREC topic were used to construct the corresponding query, each of our test queries was relatively specific. A non-relevant document (e.g., AP890524-0295) was ranked within the top 50 positions by the baseline system because of certain degree of similarity between the query and the content of the document. However, there was a large granularity gap between the specific query and the document. When we examined the content of document AP890524-0295, we found that some specific terms such as "japan", "nikkei", and "tokyo" captured in the automatically generated ontology were actually missing in the document. When we applied our granularity-based ranking mechanism (e.g., Eq. 7) to re-rank the documents, it was not difficult to find that the aggregated score of the same document became much smaller. The reason was that a large value (i.e., the amount of granularity gap) was subtracted from the document similarity score because of the granularity mismatch between the query and the document. As a result, the relative rank of document AP890524-0295 was lowered to outside the top 100 positions. On the other hand, there was little granularity gap between a relevant document such as AP890109-0325 and the query (e.g., both the document and the query were specifically about "Japanese Stock Market Trend"). Therefore, the aggregated document score of AP890109-0325 is more or less the same as its similarity score. Since some other non-specific documents were push down to the lower rank positions, the relative rank of AP890109-0325 became even higher after a granularity-based re-ranking process. As a result, the average precision is improved after applying our aggregated document ranking mechanism.

User-Based Evaluation

The aims of the user-oriented experiments are two-folds. Firstly, we examined how closely our granularity-based document ranking function approximate human's rankings with respect to the same set of documents and queries. Secondly, we compared users' perceived relevance of the top-ranked documents returned by our granular IR system and a well-known Internet search engine respectively. There were 30 human subjects (undergraduate students of various disciplines) involved in these experiments. The reason of recruiting students from various disciplines rather than just from a medical school is that we want to compare the document ranking behavior of average people with that of our system. Every subject was randomly selected and voluntarily participated in these experiments. A briefing session of 20 minutes was conducted to explain the experimental procedures and the operation of the granular IR system before the actual experiments began. There was no time limitation for the subjects to conduct information search tasks and fill in the questionnaires.

The Document Ranking Tests

As large IR benchmark collections such as the OHSUMED corpus may impose excessive cognitive load to human subjects, we developed a small collection of short documents (e.g., snippets) for the first user-oriented experiment. In this experiment, we did not explicitly mention referential or terminological specificity because human may not differentiate these concepts consciously. We employed a general Web information resource such as ask.com to extract terminological general snippets related to some diseases. In addition, we utilized PubMed to extract terminological specific snippets related to the same set of diseases. PubMed is a service of the U.S. National Library of Medicine which contains over 17 million medical citations. We employed the Spearman rank-order correlation coefficient r_s (Gan et. al. 2007) to compare the document rankings generated by human and that produced by our granular IR system. The Spearman rank-order correlation coefficient is a widely used correlation analysis method for ordinal data and it is defined by:

$$r_s = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2\right)}{n(n^2-1)} \tag{8}$$

where *n* is the number of ranks for comparison, and d_i is the difference between two corresponding ranks. We computed the Spearman rank-order correlation coefficient between every pair of human ranking and system ranking, and then we calculated the mean coefficient value for each ranking test.

For each ranking test, a subject was asked to rank three snippets with respect to their specificity to the definition of a disease. In particular, jargons (e.g., MeSH concepts) were used in the first snippet, whereas layman words were used in the second one. The third snippet referred to the definition of a financial term only (the most general or irrelevant snippet with respect to the query). The following is an example of our granularity test:

A: Bluetongue is a reovirus infection, chiefly of sheep, characterized by a swollen blue tongue, catarrhal inflammation of upper respiratory and gastrointestinal tracts, and often by inflammation of sensitive laminae of the feet and coronet. (terminologically specific – not shown to subjects)

B: Bluetongue is an insect transmitted, viral disease of sheep. However, it causes very mild, self-limiting infections with only minor clinical consequences for cattle and goats. (terminologically general - not shown to subjects)

C: A hedge fund is a private investment fund charging a performance fee and is typically reserved to a limited range of qualified investors. (both terminological and referential general - not shown to subjects)

Ten diseases were randomly chosen to construct the ranking tests. For the granular IR system, a test query was constructed based on the template "the definition of <name of a disease>", whereas the name of a particular disease was inserted into the query template. Instead of automatically estimating query granularity, a default query granularity (e.g., 1.0 for the highest specificity) was specified for the system. The experimental results are shown in Table 3. The overall mean Spearman correlation value is 0.94498 which is close to the upper bound of 1. Therefore, we believe that the document ranking mechanism of the granular IR system can reasonably imitate the document ranking processes of human information seekers.

Table 3. The Document Ranking Tests

Test	Average Spearman Correlation
African Horse Sickness	0.9833
Bluetongue	0.9167
Phlebotomus Fever	0.9333
Rift Valley Fever	0.9833
Yellow Fever	0.9833
Lung Abscess	0.9833
Pharyngitis	0.9333
Tracheitis	0.8833
Whooping Cough	0.9167
Avian influenza	0.9333
Mean	0.9450
STD	0.036

The Perceived Relevance Tests

To compare the IR effectiveness between our granular IR system and a popular Internet search engine for domain specific search, we randomly and evenly divided the human subjects into two groups; the experimental group used the granular IR system for domain specific search tasks, whereas the control group employed the Google search engine to carry out the same tasks. In addition, there were two types of domain specific search tasks for each group. Each subject needed to conduct four disease related search tasks for each type of search. The fact-finding search type (Niu and Winter 2006) employed a query template "Causes of <disease name>", and the exploratory search type was based on the query template "<disease name> related diseases". The place-holder in the template was replaced by the actual name of a disease when the experiment was conducted. Once a query was issued, subjects were not allowed to revise the query, and they were allowed to review only the top ten documents (i.e., the first page) returned by a system. After reviewing the details of each Web page pointed by the hyperlinks contained in the top result set, the subjects were requested to rate the overall "relevance" of the top ten documents based on a 5-point semantic differential scale from "highly relevant (5)", "relevant (4)", "mixed relevance (3)", "non-relevant (2)", to "highly non-relevant (1)".

Table 4. The Perceived Relevance Tests

	Granular IR System		Google	
	Mean	STD	Mean	STD
Fact-Finding Search	4.3333	0.8997	4.0667	0.9612
Exploratory Search	4.4667	0.6394	3.5333	1.1255

The independent variables are IR systems and types of search tasks, and the dependent variable is users' perceived relevance of the top-ranked Web documents. Since the average perceived relevance score of four IR tasks pertaining to a particular search type was computed before data analysis began, it was basically a 2 by 2 factorial design. For the granular IR system, manual specification of query granularity was used. In particular, subjects were told to specify a high query specificity using the granularity control bar (like the slider bar of Google Maps) when fact-finding type of search was invoked. For the exploratory search, subjects were instructed to specify low specificity for the queries. Our granular IR system employed the Google Search API[8] to retrieve the first 1,000 Web documents

[8] http://code.google.com/apis/soapsearch/

and then re-ranked these documents according to the particular query granularity specified in each run. The results of this experiment are summarized in Table 4.

The two-way ANOVA indicates no significant interaction between IR system and type of search task, $F(1, 56) = 1.96$, $p = .17$, partial $\eta^2 = .03$, but significant main effect for IR system, $F(1, 56) = 6.34$, $p = .02$, partial $\eta^2 = .10$. The mean scores of perceived relevance for the granular IR system are consistently higher than that of the Google search engine for two types of search tasks. Therefore, we conclude that our granular IR system produces more relevant information in the first page of the result set when compared with that of the Google search engine. Further investigation into this experiment revealed that a combined similarity and popularity ranking as employed by Google may not always produce the most relevant results. For instance, with reference to the specific search for the "causes of African Horse Sickness (AHS)", the first document returned by Google was the facebook community page[9] about the AHS issues in general. Although this page is very popular (frequently referred to by a large facebook community), it is not relevant to our specific query. On the other hand, for exploratory information search such as "AHS related diseases", the first document returned by Google was the Oie page[10] which was related to the epidemiology and diagnosis of AHS. It is still very specific about AHS rather than the AHS related diseases. In fact, most of the links in the first result page produced by Google pointed to very specific AHS documents even though general AHS documents were required. In contrast, our granular IR system could detect such a granularity gap between the Oie page and the general AHS query according to the proposed granularity-based ranking mechanism; instead of including the Oie page in the first result page, other Web documents about the general issues of AHS (e.g., related diseases) were ranked much higher. As a result, the perceived relevance of the first result page returned by the granular IR system was higher than that of the result set returned by Google.

Conclusions and Future Work

Because of the rapid growth of heterogeneous information archived on networks of computers such as the Internet, it is increasingly more difficult for information seekers to retrieve relevant information. The widely used similarity-based and popularity-based document ranking functions can be improved to alleviate the problem of information overload. By exploiting the granular computing methodology, we design and develop a novel granular IR system to enhance domain specific search. In particular, a computational model is developed to rank documents according to the specific granularity requirements of the information seekers. Large benchmark corpora were applied to evaluate the effectiveness of the granular IR system. The initial experimental results confirm that our granular IR system outperforms a classical similarity-based IR system for the routing tasks. In addition, user-oriented experiments show that the perceived relevance of the top-ranked documents recommended by our system is higher than that of the documents returned by a well-known search engine. Our research work opens the door to the design of the next generation of Internet search engines, and it sheds light on developing more sophisticated IR systems to alleviate the problem of information overload. In the future, we will apply our granular IR system to search for information for a variety of domains by using a general ontology such as LCSH. Moreover, the optimal values of the system parameters will be sought by invoking heuristic search methods such as a genetic algorithm.

Acknowledgements

The work reported in this paper has been funded by the Start-Up grant of City University of Hong Kong (Project No.: 7200126). The authors would like to thank the three anonymous reviewers, the associate editor, and the track chair for valuable suggestions that help a lot in improving the quality of the paper.

References

Bargiela, A. and Pedrycz, W. "Toward a Theory of Granular Computing for Human-Centered Information Processing", *IEEE Transactions on Fuzzy Systems* (16:2), 2008, pp. 320-330.

Buyukkokten, O., Kaljuvee, O., Garcia-Molina, H., Paepcke, A., and Winograd, T. "Efficient Web browsing on handheld devices using page and form Summarization", *ACM Transactions on Information Systems* (20:1), 2002, pp. 82-115.

[9] http://apps.facebook.com/causes/8336?facebook_url=true as accessed on 14 April 2008

[10] http://www.oie.int/eng/maladies/fiches/A_A110.HTM as accessed on 14 April 2008

Ferizis, G., and Bailey, P. "Towards practical genre classification of web documents", in *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland, May 2006, pp. 1013-1014.

Gan, G., Ma, C., and Wu, J. Data Clustering: Theory, Algorithms, and Applications. SIAM, 2007.

Granka, L., Joachims, T., and Gay, G. "Eye-tracking analysis of user behavior in www-search", in *Proceedings of the 27th ACM SIGIR International Conference on Research and Development in Information Retrieval*, Sheffield, UK, July 2004, pp. 478-479.

Griffiths, T., Steyvers, M., and Firl, A. "Google and the Mind: Predicting Fluency With PageRank", *Psychological Science* (18:12), 2007, pp. 1069–1076.

Gruber, T.R. "A translation approach to portable ontology specifications", *Knowledge Acquisition*, (5:2), 1993, pp. 199–220.

Haveliwala, T.H. "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search", *IEEE Transactions on Knowledge and Data Engineering* (15:4), 2003, pp. 784-796.

Hersh, W., Buckley, C., Leone, T.J., and Hickam, D. "Ohsumed: an interactive retrieval evaluation and new large test collection for research", in *Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 3-6 July 1994, pp. 192 - 201.

Hevner, A., March, S., Park, J., Ram, S. "Design science in information systems research", *MIS Quarterly* (28:1), 2004, pp. 75-105.

Hull, D. "The TREC-7 Filtering Track: Description and Analysis", *Proceedings of the 7th Text Retrieval Conference*, Gaithersburg, MD, USA, 1998, pp. 33-56.

Lau, R.Y.K., Bruza, P.D., and Song, D. "Towards a Belief Revision Based Adaptive and Context Sensitive Information Retrieval System", *ACM Transactions on Information Systems* (26:2), 2008. To appear.

Lau, R.Y.K., Li, Y., and Xu, Y. "Mining Fuzzy Domain Ontology from Textual Databases", in *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence*, Silicon Valley, USA, 2007, pp. 156-162.

Lawrence, S. "Context in Web Search", *IEEE Data Engineering Bulletin* (23:3), 2000, pp. 25-32.

Mowshowitz, A. and Kawaguchi, A. "Bias on the Web", *Communications of the ACM* (45:9), 2002, pp. 56–60.

Niu, Y. and Winter, S. "The Effect of Task Type and Information Format on Web Searching Performance", in *Proceedings of the Twenty-Seventh International Conference on Information Systems*, Haseman, W.D. (Ed.), Milwaukee, Wisconsin, December 2006, pp. 791-802.

Page, L., Brin, S., Motwani, R. and Andwinograd, T. "The PageRank citation ranking: Bringing order to the web", *Technical Report*, Stanford Digital Library Technologies Project, 1998.

Ponte, J.M. and Croft, W.B. "A Language Modeling Approach to Information Retrieval", in *Proceedings of the 21st ACM SIGIR International Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998, pp. 275-281.

Porter, M. "An algorithm for suffix stripping", *Program*, 1980, (14:3), pp.130–137.

Roussinov, D. G., and Chen, H. "Information Navigation on the Web by Clustering and Summarizing Query Results", *Information Processing and Management* (37:6), November 2001, pp. 789 – 816.

Salton, G., Wong, A., and Yang, C.S. "A vector space model for automatic indexing", *Communications of the ACM* (18:11), November 1975, pp. 613-620.

Salton, G. "Full text information processing using the smart system", *Database Engineering Bulletin* (13:1), March 1990, pp. 2–9.

Salton, G. "Developments in automatic text retrieval", *Science* (253:5023), 1991, pp. 974-980.

Van Rijsbergen, C.J. Information Retrieval. London, UK: Butterworths, 1979.

Yan, X., Song, D., and Li, S. "Concept-based document readability in domain specific information retrieval", in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, VA, USA, 2006, pp. 540-549.

Yao, Y.Y. "Information retrieval support systems", in *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, 2002, pp. 773-778.

Yao, Y.Y. "Perspectives of granular computing", in *Proceedings of the 2005 IEEE International Conference on Granular Computing*, 2005, pp. 85-90.

Yao, J.T. "Information granulation and granular relationships", in *Proceedings of the 2005 IEEE International Conference on Granular Computing*, 2005, pp. 326-329.

Zadeh, L.A. "Fuzzy sets and information granularity", in *Advances in Fuzzy Set Theory and Applications*, M. Gupta, R.K. Ragade, R.R. Yager (eds), 1979, North-Holland Publishing Company, pp. 3-18.

Zhou, X., Wu, S.T., Li, Y., Xu, Y., Lau, R.Y.K., and Bruza, P.D. "Utilizing Search Intent in Topic Ontology-Based User Profile for Web Mining", in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Hong Kong, China, 2006, pp. 558-564.