Fuzzy Ontology Mining and Semantic Information Granulation for Effective Information Retrieval Decision Making

Raymond Y.K. Lau [1] , Chapmann C.L. Lai [1] , Yuefeng Li [2]

[1] *Department of Information Systems, City University of Hong Kong,*
Tat Chee Avenue, Kowloon, Hong Kong
E-mail: {raylau, chunllai}@cityu.edu.hk

[2] *School of Information Technology, Queensland University of Technology,*
GPO Box 2434, Brisbane, Qld 4001, Australia
E-mail: y.li@qut.edu.au

Accepted: 29-11-2009
Received: 09-09-2010

Abstract

The notion of semantic information granulation is explored to estimate the information specificity or generality of documents. Basically, a document is considered more specific than another document if it contains more cohesive domain-specific terminologies than that of the other one. We believe that the dimension of semantic granularity is an important supplement to the existing similarity-based and popularity-based measures for building effective document ranking functions. The main contributions of this paper is the illustration of the design and development of a fuzzy ontology based granular information retrieval (IR) system to improve the effectiveness of IR decision making for various domains. Based on the notion of semantic information granulation, a novel computational model is developed to estimate the semantic granularity of documents; these documents can then be ranked according to the information seekers' specific semantic granularity requirements. One main component of the proposed computational model is the fuzzy ontology mining mechanism which can automatically build domain-specific ontology for the estimation of semantic granularity of documents. Our TREC-based experiment reveals that the proposed fuzzy ontology based granular IR system outperforms a classical vector space based IR system in domain specific IR. Our research work opens the door to the applications of granular computing and fuzzy ontology mining methods to enhance domain specific IR decision making.

Keywords: Text Mining, Fuzzy Ontology, Fuzzy Subsumption, Information Granulation, Granular Computing, Information Retrieval.

1. Introduction

Classical similarity-based document ranking functions [22,23], and the recent popularity-based ranking algorithms [16] have been applied to developed IR systems such as Internet search engines. 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 could be weak for newly created Web documents [14]. In this paper, we examine if another dimension can be used to supplement the existing similarity-based and popularity-based dimensions so as to develop a more effective document ranking function. In fact, recent advances in Granular Computing [1,31,32] sheds

Published by Atlantis Press

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 [1,13,19,21,32]. In granular computing, information granulation refers to the computational processes of generating and presenting levels of granularity of information to facilitate problem solving [30,34,17]. However, when researchers refer to information granularity in the discipline of granular computing, they often mean the "structural granularity", that is, the structural abstractions of information items such as the coarse level of a document containing the finer levels of sections, chapters, paragraphs, sentences, and so on. In the context of IR, we try to apply the concept of information granulation to design a granular IR system which can estimate the "semantic granularity" of documents (e.g., general vs. specific documents) and rank these documents with respect to information seekers' specific granularity requirements.

The semantic granularity of a document refers to the levels of semantic details (i.e., the specificity) of information contained in the document [9]. To objectively estimate the semantic granularity of a document, it is possible to refer to a domain ontology such as the Medical Subject Headings (MeSH)[a] to determine if a document contains specific terminologies or general concepts. It is generally accepted that ontology refers to a formal specification of conceptualization [3]. Since specificity is the antonym of generality, we will measure a document's semantic granularity (an attribute) in terms of document specificity (attribute value) throughout this paper. As we only focus on the semantic granularity rather than the structural granularity of documents, we will loosely use the term granularity to refer to semantic granularity for the rest of this paper.

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 the granularity of a document is subject to human interpretations in some cases. For instance, the term "granular computing" may be considered specific for an undergraduate student, while it may be considered too general for researchers in the field of granularity computing. One of the main contributions of this paper is the development of a computational model for granular information retrieval. Such a computational model supports the objective estimation of the granularity of documents by consulting domain ontology, and it can also deal with the subjectivity in information granularity by taking into account the individual user's granularity requirement. As comprehensive domain ontology may not be readily available for arbitrary domains, the second main contribution of this paper is the illustration of how to apply a fuzzy ontology discovery method to enhance the effectiveness of the proposed granular IR system.

The remainder of the paper is organized as follows. Section 2 highlights previous research in the related area and compare these research work with ours. Section 3 describes the general architecture of the proposed granular IR system. Our novel fuzzy ontology discovery method is then illustrated in Section 4. The computational details for the estimation of document or query granularity are then given in Section 5. Section 6 describes TREC-AP collection based evaluation of the proposed granular IR system. Finally, we offer concluding remarks and describe future directions of our research work.

2. Related Research

Yao [31] 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 [31]. However, the idea of applying the granular computing methodology to IR remains as a concep-

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

tual 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. Based on the conceptual ideas presented in [31], we explore term space granulation and apply this concept to construct a computational model to estimate the granularity of documents and queries.

It was argued that the readability of a document could be assessed quantitatively and objectively [29]. 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. 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. [35] 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.

The FOGA framework for fuzzy ontology generation has been proposed [27]. The FOGA framework consists of fuzzy formal concept analysis, fuzzy conceptual clustering, fuzzy ontology generation, and semantic representation conversion. Essentially, the FOGA method extends the formal concept analysis approach, which has also been applied to ontology discovery, with the notions of fuzzy sets. Our proposed ontology discovery method is based on previous work in computational linguistic and with the computational mechanism built based on the concept of fuzzy relations.

An ontology mining technique was proposed to extract patterns representing users' information needs [12]. The ontology mining method consists of two parts: the top backbone and the base backbone. The former represents the relations between compound classes of the ontology. The latter indicates the linkage between primitive classes and compound classes. The Dempster-Shafer theory of evidence model was applied to extract the relations among classes. The strength of the ontology mining method is that it can effectively synthesize taxonomic relations and non-taxonomic relation in a single ontology model. In addition, a novel method was proposed to capture the evolving patterns in order to refine the initially discovered ontology. Finally, a formal model was developed to assess the relevance of the discovered ontology with respect to the user's information needs. The research work presented in this paper focuses on fuzzy domain ontology discovery rather than the discovery of crisp ontology representing users' information needs.

A fuzzy ontology which is an extension of the domain ontology with crisp concepts is utilized for news summarization purpose [11]. The main function of the fuzzy inference mechanism is to generate the membership degrees (classification) for each event with respect to some pre-defined concepts. The standard triangular membership function is used for the classification purpose. The method discussed in this paper is a fully automatic fuzzy domain ontology discovery approach. There is no pre-defined fuzzy concepts and taxonomy of concepts, instead our text mining method will automatically discover such concepts and generate the taxonomy relations.

Sanderson and Croft [24] proposed a document-based subsumption induction method to automatically derive a hierarchy of terms from a corpus. In particular, the subsumption relations among terms are developed based on the co-occurrence of terms in the documents of a corpus. For example, term t_1 is considered more specific than another term t_2 if the appearance of t_1 in a document implies the appearance of t_2 in the same document but not vice versa. They adopted an artificial threshold such as $Pr(t_2|t_1) \geqslant 0.8$ as a fixed cut-off to determine the specificity relation between t_1 and t_2. Our concept mining method differs from their work in that we are dealing with the more challenging task of concept hierarchy mining rather than term relationship extraction. In addition, our method extends their computational method in that the co-occurrence of terms is derived based on a moving text window rather than the whole document to reduce the chance of generating noisy subsumption relations.

3. General System Architecture

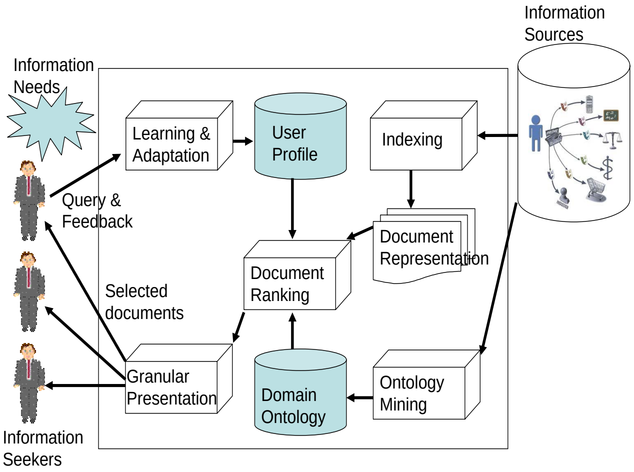

Fig. 1. The System Architecture of the Granular IR System

The general system architecture of our granular IR system is depicted in Figure. 1. 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. To alleviate the problem of lack of good quality domain ontology for certain domains, the ontology mining mechanism can automatically discover domain ontology based on some information sources. Ontology discovery is invoked as a background task, and so it will not affect the online IR performance. Our prototype system[b] was developed using Java (J2SE v 1.4.2), Java Server Pages (JSP) 2.1, and Servlet 2.5.

4. Automatic Fuzzy Domain Ontology Discovery

4.1. A Formal Model for Fuzzy Domain Ontology

Since any IR processes involve uncertainty [8], an uncertainty management mechanism is required in our ontology mining method. The notions of Fuzzy set and Fuzzy Relation are effective to represent knowledge with uncertainty [33]. Therefore, a fuzzy ontology rather than a crisp ontology is employed in our granular IR system. Our proposed fuzzy domain ontology is formally defined by:

Definition 1. [Fuzzy Set] A fuzzy set $\mathscr{F}$ consists of a set of objects drawn from a domain X and the

[b] http://quantum.is.cityu.edu.hk/ong/ongui.jsp

membership of each object x_i in $\mathscr{F}$ is defined by a membership function $\mu_{\mathscr{F}} : X \mapsto [0,1]$. For the special case of a crisp set, the crisp membership function has the mapping $\mu_{\mathscr{F}} : X \mapsto \{0,1\}$.

Definition 2. [Fuzzy Relation] A fuzzy relation R_{XY} is defined as the fuzzy set $\mathscr{R}$ on a domain $X \times Y$ where X and Y are two crisp sets. The membership of each object (x_i, y_i) in $\mathscr{R}$ is defined by a membership function $\mu_{\mathscr{R}} : X \times Y \mapsto [0,1]$.

Definition 3. [Fuzzy Ontology] A fuzzy ontology is a 6-tuple $Ont = \langle X, A, C, R_{XC}, R_{AC}, R_{CC} \rangle$, where X is a set of objects, A is the set of attributes describing the objects, and C is a set of concepts (classes). The fuzzy relation $R_{XC} : X \times C \mapsto [0,1]$ assigns a membership to the pair (x_i, c_i) for all $x_i \in X, c_i \in C$, the fuzzy relation $R_{AC} : A \times C \mapsto [0,1]$ defines the mapping from the set of attributes A to the set of concepts C, and the fuzzy relation $R_{CC} : C \times C \mapsto [0,1]$ defines the strength of the sub-class/super-class relationships among the set of concepts C.

Based on the idea of formal concept analysis [2], X is the *extent* of the concepts C, and A is the *intent* which defines the properties of C. According to the idea of subsumption, the sub-concept/super-concept relation (R_{CC}) can be defined by:

Definition 4. [Fuzzy Subsumption] With respect to an arbitrary α-cut level, a concept $c_x \in C$ is the sub-concept of another super-concept $c_y \in C$ if and only if $\forall a_i \in \{z \in A | \mu_{R_{AC}}(z, c_y) \geqslant \alpha\}$, $\mu_{R_{AC}}(a_i, c_x) \geqslant \alpha$. Alternatively, from an extensional perspective, a concept $c_x \in C$ is the sub-concept of another super-concept $c_y \in C$ if and only if $\forall x_i \in \{z \in X | \mu_{R_{XC}}(z, c_x) \geqslant \alpha\}$, $\mu_{R_{XC}}(x_i, c_y) \geqslant \alpha$ with respect to an arbitrary α-cut level.

Definition 4 can be explained as follows: if the membership of every attribute $a_i \in A$ for the concept $c_y \in C$ is greater than or equal to a certain threshold α, the membership of the corresponding attribute a_i for the concept $c_x \in C$ is also greater than or equal to α, then the concept c_x is the sub-concept of c_y. As can be seen, the crisp subsumption relation is only a special case of the generalized fuzzy subsumption relation in that the threshold value $\alpha = 1$ is established for the crisp case. In other words, if it is true that every attribute $a_i \in A$ characterizing the concept c_y implies that it also characterizes the concept c_x, the concept c_x is the sub-concept of c_y.

4.2. Context-Sensitive Text Mining for Concept Discovery

In the field of IR, the notion of *context vectors* [5,25] has been proposed to construct computer-based representations of concepts (i.e., linguistic class). In this approach, a concept is represented by a vector of words (features) and their numerical weights. The weight of a word indicates the extent to which the particular word is *associated* with the underlying concept. Figure. 2 shows that the concept “acquisition” is represented by the feature terms such as “merger”, “company”, “arbitrage”, “takeover”, “buyout”, etc. Indeed, this is a real example extracted from the TREC-AP Topic description file by applying our ontology discovery algorithm. All the terms are stemmed by our program as shown in Figure. 2. The context vector of “acquisition” is represented as follows:

Concept: acquisition
Context Vector:
$\langle (merger, 0.675), (compani, 0.675), (arbitrag, 0.589),$
$(takeover, 0.507), (buyout, 0.416), \ldots \rangle$

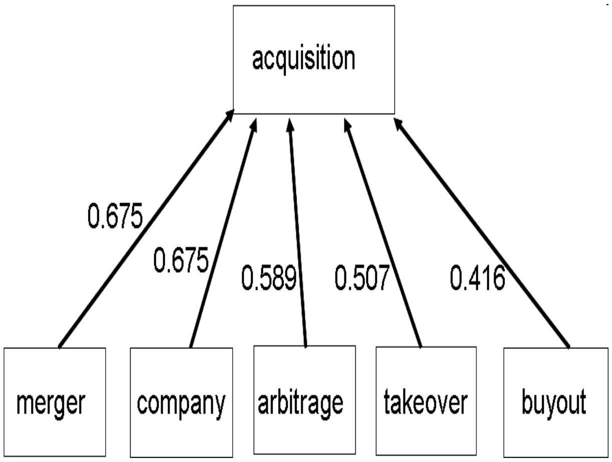

Fig. 2. The Representation of the Concept “Acquisition”

A context vector can be seen as a point in a multi-dimensional geometric information space with each dimension representing a property term. A linguistic concept such as “acquisition” can be taken as a class (set) with respect to the fuzzy sets framework. A feature word such as “merger” can then be treated as an attribute describing the concept to a certain degree (i.e., $\mu_{R_{AC}}(merger, acquisition) = 0.675$).

Our context-sensitive text mining method is first applied to extract concepts from a textual corpus [6]. In particular, a virtual window is moved from left to right among the tokens in each document of a corpus to take into account the proximity among terms. According to previous studies, a text window of 5 to 10 terms is effective [7,18], and so we adopt this range as the basis to perform our windowing process. Standard association measure such as Mutual Information (MI) [26] is then applied to measure the association between a concept and its underlying terms:

$$MI(t_i,t_j) = \log_2 \frac{Pr(t_i,t_j)}{Pr(t_i)Pr(t_j)} \tag{1}$$

where $MI(t_i,t_j)$ is the mutual information between term t_i and term t_j. $Pr(t_i,t_j)$ is the joint probability that both terms appear in a text window, and $Pr(t_i)$ is the probability that a term t_i appears in a text window. The probability $Pr(t_i)$ is estimated based on $\frac{|w_t|}{|w|}$ where $|w_t|$ is the number of windows containing the term t and $|w|$ is the total number of windows constructed from a corpus. Similarly, $Pr(t_i,t_j)$ is the fraction of the number of windows containing both terms out of the total number of windows.

We develop *Balanced Mutual Information* (BMI) [7,10] to compute the degree of association among tokens. This method considers both term presence and term absence as the evidence of the implicit term relationships.

$$\begin{aligned}\mu_{c_i}(t_j) \quad & \approx BMI(t_i,t_j) \\ & = \beta \times [Pr(t_i,t_j)\log_2(\tfrac{Pr(t_i,t_j)+1}{Pr(t_i)Pr(t_j)}) + \\ & Pr(\neg t_i,\neg t_j)\log_2(\tfrac{Pr(\neg t_i,\neg t_j)+1}{Pr(\neg t_i)Pr(\neg t_j)})] - \\ & (1-\beta) \times [Pr(t_i,\neg t_j)\log_2(\tfrac{Pr(t_i,\neg t_j)+1}{Pr(t_i)Pr(\neg t_j)}) + \\ & Pr(\neg t_i,t_j)\log_2(\tfrac{Pr(\neg t_i,t_j)+1}{Pr(\neg t_i)Pr(t_j)})]\end{aligned} \tag{2}$$

where $\mu_{c_i}(t_j)$ is the membership function to estimate the degree of a term $t_j \in A$ belonging to a concept $c_i \in C$. $\mu_{c_i}(t_j)$ is the computational mechanism for the relation R_{AC} defined in the fuzzy domain ontology $Ont = \langle X, A, C, R_{XC}, R_{AC}, R_{CC} \rangle$. The membership function $\mu_{c_i}(t_j)$ is indeed approximated by the BMI score. The weight factor $\beta > 0.5$ is used to control the relative importance of two kinds of evidence (positive and negative).

To improve computational efficiency and filter noisy relations, only certain linguistic pattern (e.g., Noun Noun, and Adjective Noun) will be considered. After the linear normalization process (i.e., $\forall_{c_i \in C, t_j \in X} \mu_{c_i}(t_j) \in [0,1]$), concept vectors [5,25] can be built. Essentially, the MI score between a concept and an underlying feature word is treated as the membership of the word (attribute) characterizing the concept. An α-cut is applied to discard terms from the potential concept if their membership values are below the threshold α. For instance, the concept vector for “commercial bank” is represented by $c_x = \langle (merger, 0.675)$, $(compani, 0.675)$, $(arbitrag, 0.589)$, $(takeover, 0.507) \rangle$ after applying a cut of $\alpha = 0.5$.

4.3. *Concept Pruning*

To further filter the noisy concepts, we adopt the TFIDF [23] like heuristic to perform the filtering process. Similar approach has also been used in ontology learning [15]. For example, if a concept is significant for a particular domain, it will appear more frequently in that domain when compared with its appearance in other domains. The following measure is used to compute the relevance score of a concept:

$$Rel(c_i,D_j) = \frac{Dom(c_i,D_j)}{\sum_{k=1}^{n} Dom(c_i,D_k)} \tag{3}$$

where $Rel(c_i,D_j)$ is the relevance score of a concept c_i in the domain D_j. The term $Dom(c_i,D_j)$ is the domain frequency of the concept c_i (i.e., number of documents containing the concept divided by the total number of documents in the corpus). The higher the value of $Rel(c_i,D_j)$, the more relevant the concept is for domain D_j. Based on empirical testing, we can estimate a threshold ϖ for a particular domain. Only the concepts with relevance scores greater than the threshold will be selected.

4.4. *Fuzzy Relation Discovery*

The final stage towards our ontology discovery method is fuzzy taxonomy generation based on subsumption relations among extracted concepts. $Spec(c_x, c_y)$ denotes that concept c_x is a specialization (sub-class) of another concept c_y. The degree of such a specialization is derived by:

$$\mu_{C\times C}(c_x, c_y) \approx Spec(c_x, c_y) = \frac{\sum_{t_x \in c_x, t_y \in c_y, t_x = t_y} \mu_{c_x}(t_x) \otimes \mu_{c_y}(t_y)}{\sum_{t_x \in c_x} \mu_{c_x}(t_x)} \quad (4)$$

where $\otimes$ is a fuzzy conjunction operator which is equivalent to the min function. The above formula states that the degree of subsumption (specificity) of c_x to c_y is based on the ratio of the sum of the minimal membership values of the common terms belonging to the two concepts to the sum of the membership values of terms in the concept c_x. The range of the $Spec(c_x, c_y)$ function falls in the unit interval $[0, 1]$ and the subsumption relation is asymmetric. When the taxonomy is built, we only select the subsumption relations such that $Spec(c_x, c_y) > Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ where λ is a threshold to distinguish significant subsumption relations. The parameter λ is estimated based on empirical tests. In addition, a pruning step is introduced such that the redundant taxonomy relations are removed. If the membership of a relation $\mu_{C\times C}(c_1, c_2) \leqslant \min(\{\mu_{C\times C}(c_1, c_i), \ldots, \mu_{C\times C}(c_i, c_2)\})$, where $c_1, c_i, \ldots, c_2$ form a path P from c_1 to c_2, the relation $R_{(}c_1, c_2)$ is removed because it can be derived from other stronger taxonomy relations in the ontology. The details of the fuzzy domain ontology mining algorithm can be found at [10].

5. A Computational Model for Estimating Semantic Granularity

5.1. *Semantic Information Granulation*

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. With reference to Figure. 3 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. On the other hand, 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. 3, 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 dimensions, that is, *terminological specificity* and *referential specificity*.

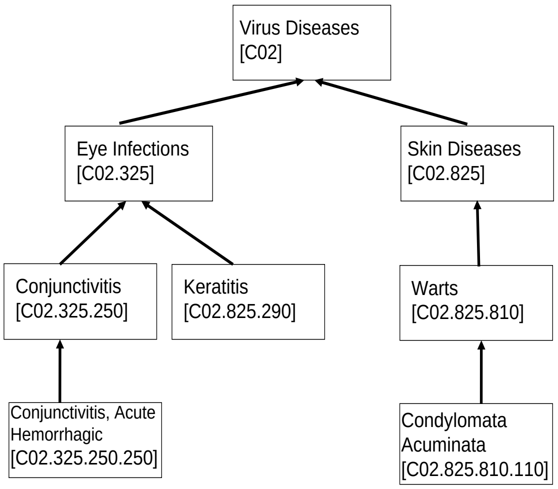

Fig. 3. A Segment of the MeSH Domain Ontology

5.2. Computing Document and Query Granularity

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. 3, the more low-level terminologies appear in a document, the more specific the document will be. Eq.(5) is proposed to measure the terminological specificity of a document by computing the average depths of all the document terms with reference to a domain ontology. 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.

$$TS(d) = \frac{\sum_{t \in MC} depth(t)}{max(ran(depth)) \times |MC|} \tag{5}$$

where MC refers to the set of matching terminologies found in a document d and an ontology, and $|MC|$ is the cardinality of the set MC. The function $depth(t)$ returns the depth of the terminology t with respect to an ontology. 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.

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 referential specificity of two terms can be measured based on the information specificity (information content) of their least common subsumer [20]:

$$RS(d) = \frac{2\sum_{t_i \in MC, t_j \in MC, i \neq j} sim(t_i, t_j)}{|MC| \times (|MC| - 1)} \tag{6}$$

$$sim(t_i, t_j) = \max_{cs \in S(t_i, t_j)} -\log_2 Pr(cs) \tag{7}$$

where the function $sim(t_i, t_j)$ returns the "semantic similarity" of two terminologies t_i and t_j based on the information content of their least common subsumers found in an ontology. $S(t_i, t_j)$ represents the set of common subsumers cs of t_i and t_j. The factor $\frac{|MC| \times (|MC|-1)}{2}$ returns the number of distinct terminology pairs constructed from the set MC. The function $RS(d)$ basically tries to measure the average semantic similarity among the pairs of matching domain terminologies found in the document d. The probability $Pr(cs)$ of a concept cs is estimated based on:

$$Pr(cs) = \frac{TF(cs)}{N} \tag{8}$$

$$TF(cs) = \sum_{x \in Subsumed(cs)} count(x) \tag{9}$$

where $TF(cs)$ represents the frequency of the concept cs estimated according to the structure of an ontology and the populated concept frequencies derived from a chosen document collection. The term N represents the sum of all concept frequencies with reference to the ontology and the chosen document collection. When the frequency of cs is estimated, any concepts subsumed by cs is also taken into account. The set $Subsumed(cs)$ represents all the concepts subsumed by cs according to the ontology. The function $count(x)$ simply returns the occurrence frequency of the concept x with reference to the ontology and the chosen document collection. For the experiment reported in this paper, we adopted the TREC-AP collection to estimate concept probability. The TREC-AP collection consists of $1,226,194$ unique terms.

By taking into account both terminological specificity and referential specificity, the specificity of a document can be estimated according to Eq.(10). 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 Eq.(11). The weight factor $\varphi_d \in [0,1]$ controls the relative importance of terminological specificity and referential specificity in estimating the overall document specificity. Similarly, the weight factor $\varphi_q \in [0,1]$ 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.(11) 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.

$$DS(d) = \varphi_d \times TS(d) + (1 - \varphi_d) \times RS(d) \qquad (10)$$

$$QS(q) = \varphi_q \times TS(q) + (1 - \varphi_q) \times RS(q) \qquad (11)$$

To implement a holistic ranking function, our granular IR system can re-rank documents after applying a similarity or popularity based ranking function. 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 φ_G controls the relative weight of the granularity factor when documents are ranked. The function $SimPop(d,q)$ represents the combined similarity and popularity based document ranking function; it is assumed that such a ranking function has been made available inside or outside (e.g., from Internet search engines) our proposed granular IR system.

$$GScore(d,q) = SimPop(d,q) - \varphi_G \times |DS(d) - QS(q)| \qquad (12)$$

6. Experiments and Results

Our experimental procedure was based on the routing task employed in the TREC forum [28]. 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 [28]. 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. In particular, we employed the TREC evaluation package available at Cornell University to compute all the performance data. We used the TREC-AP collection which comprises the Associated Press (AP) newswires covering the period from 1988 to 1990 with a total number of 242,892 documents (with the removal of 26 documents containing stop words only) and the average document length of 450.6 words [4].

Table 1. Results of the TREC-AP Benchmark Test

Recall	Baseline System		Granular IR System		t-statistics	p values
	Mean	STD	Mean	STD	$df(9)$	
0	0.5223	0.1001	0.5978	0.1013	5.285	$< .01^{**}$
0.1	0.3035	0.1730	0.3781	0.1826	4.911	$< .01^{**}$
0.2	0.2325	0.1888	0.2862	0.2019	4.591	$< .01^{**}$
0.3	0.1994	0.1706	0.2432	0.2115	3.716	$< .01^{**}$
0.4	0.1687	0.1528	0.2168	0.2109	3.354	$< .01^{**}$
0.5	0.1240	0.1115	0.1872	0.2033	2.719	$= .01^{**}$
0.6	0.0869	0.0791	0.1425	0.1384	2.443	$< .05^{*}$
0.7	0.0644	0.0743	0.1063	0.1052	2.069	$< .05^{*}$
0.8	0.0390	0.0759	0.0608	0.0554	1.081	$= .15$
0.9	0.0103	0.0235	0.0223	0.0311	1.179	$= .13$
1	0.0047	0.0026	0.0158	0.0250	1.536	$= .08$
MAP	0.1519		0.1782			
Δ%			17.31%			

A baseline system was developed based on the classical vector space model [23]. 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 the aggregated document ranking function Eq.(12) to rank documents. The query specificity of each TREC-AP query was computed according to Eq.(11). For all the experiments reported in this paper, the parameters $\varphi_d = \varphi_q = 0.41$ and $\varphi_G = 0.83$ were used. These system parameters were estimated based on the pilot tests which involved a subset of the TREC-AP test queries.

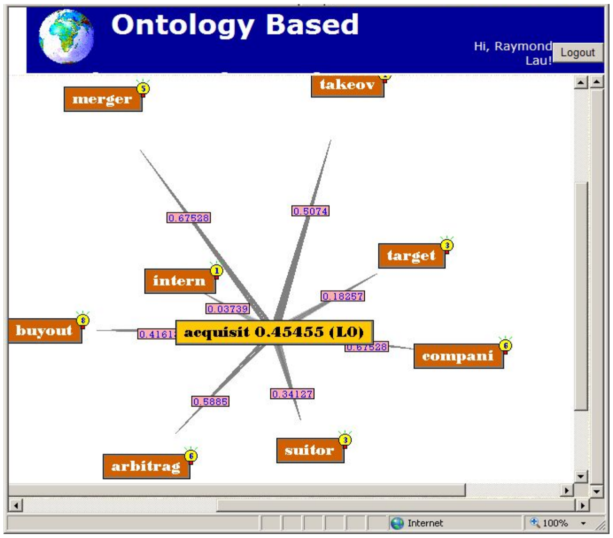

Fig. 4. The First Level Concepts of the "Acquisitions" Ontology

We randomly selected ten TREC-AP 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 discovery method to automatically generate a domain ontology based on the full-text description of the topic. Figure. 4 shows the first level concepts of the "Acquisitions" ontology after applying our fuzzy domain ontology mining algorithm to the TREC Topic 002; there are lower levels concepts which are not expanded in the diagram for better readability reason. The performance data as generated by the TREC evaluation package is tabulated in Table. 1. 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 system and the baseline IR system respectively.

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 17.31% overall improvement, and such an improvement is shown to be statistically significant. 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. The performance improvement of the granular IR system over the baseline system can be explained with reference to Figure. 4. For instance, the concepts "merger", "arbitrage", "takeover", "buyout", and so on captured in the system generated fuzzy domain ontology were used to estimate the *specificity* of the TREC-AP documents when the topic (query) "acquisitions" was processed. This extra semantic information helped a lot in determining if certain documents were specific about the topic "acquisitions" or not, and accordingly the rank of these documents was adjusted.

7. Conclusions and Future Work

By exploiting the granular computing methodology, we design and develop a novel granular IR system to enhance document retrieval decision making for specific domains. In particular, the notion of semantic granularity is proposed and the corresponding computational model is developed to estimate the semantic granularity (i.e., specific vs. general) of documents and queries. One main component of our proposed computational model is the fuzzy domain ontology mining mechanism. By automatically extracting the fuzzy ontology pertaining to various domains, our system can effectively measure the semantic granularity of documents. A TREC-based benchmark corpus was applied to evaluate the effectiveness of the fuzzy ontology based granular IR system. Our experimental results show that the fuzzy ontology based granular IR system outperforms a classical similarity-based IR system for some routing tasks. Our research work opens the door to the design of the next generation of domain-specific IR systems such as domain-specific search engines. In the future, we will evaluate our granular IR system by comparing it with a baseline system which can utilize a general ontology such as the Library of Congress Subject Headings (LCSH) for IR. Moreover, the optimal values of the system parameters will be sought by invoking heuristic search methods such as a genetic algorithm. Finally, field tests will be conducted to compare the IR effectiveness between our fuzzy ontology based granular IR system and Internet search engines.

Acknowledgments

The work reported in this paper has been funded in part by City University of Hong Kong's Start-up Grant (Project No.: 7200126), Strategic Research Grant (Project No.: 7002426), and Strategic Research Grant (Project No.: 7008018).

References

1. A. Bargiela and W. Pedrycz. Toward a theory of granular computing for human-centered information processing. *IEEE Transactions on Fuzzy Systems*, 16(2):320–330, 2008.
2. P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.
3. T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
4. D. Hull. The TREC-7 Filtering Track: Description and Analysis. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the seventh Text REtrieval Conference (TREC-7)*, pages 33–56, Gaithersburg, Maryland, November 9–11 1998. NIST. Available from `http://trec.nist.gov/pubs/trec7/`.
5. Hongyan Jing and Evelyne Tzoukermann. Information retrieval based on context distance and morphology. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Language Analysis, pages 90–96, 1999.
6. R.Y.K. Lau. Context-Sensitive Text Mining and Belief Revision for Intelligent Information Retrieval on the Web. *Web Intelligence and Agent Systems An International Journal*, 1(3-4):1–22, 2003.
7. R.Y.K. Lau. Fuzzy Domain Ontology Discovery for Business Knowledge Management. *IEEE Intelligent Informatics Bulletin*, 8(1):29–41, 2007.
8. R.Y.K. Lau, P. Bruza, and D. Song. Towards a Belief Revision Based Adaptive and Context-Sensitive Information Retrieval System. *ACM Transactions on Information Systems*, 26(2):8.1–8.38, 2008.
9. R.Y.K. Lau, C.L. Lai, and Y. Li. Mining fuzzy ontology for a web-based granular information retrieval system. In Peng Wen, Yuefeng Li, Lech Polkowski, Yiyu Yao, Shusaku Tsumoto, and Guoyin Wang, editors, *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*, volume 5589 of *Lecture Notes in Computer Science*, pages 239–246, Gold Coast, Australia, July 14–16 2009. Springer.
10. R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, and J.X. Hao. Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):800–813, 2009.
11. Chang-Shing Lee, Zhi-Wei Jian, and Lin-Kai Huang. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(5):859–880, 2005.
12. Yuefeng Li and Ning Zhong. Mining ontology for automatically acquiring web user information needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.
13. Luis Martinez, Manuel J. Barranco, Luis G. Perez, and Macarena Espinilla. A knowledge based recommender system with multigranular linguistic information. *International Journal of Computational Intelligence Systems*, 1:225–236, 2008.
14. A. Mowshowitz and A. Kawaguchi. Efficient web browsing on handheld devices using page and form summarization. *Communications of the ACM*, 45(9):56–60, 2002.
15. Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31, 2003.
16. L. Page, S. Brin, R. Motwani, and T. Andwinograd. The PageRank citation ranking: Bringing order to the Web. In *Stanford Digital Library Technologies*

Project, 1998. Technical Report.

17. Witold Pedrycz. Hierarchical architectures of fuzzy models: From type-1 fuzzy sets to information granules of higher type. *International Journal of Computational Intelligence Systems*, 3:202–214, 2010.
18. Patrick Perrin and Frederick Petry. Extraction and representation of contextual information for knowledge discovery in texts. *Information Sciences*, 151:125–152, 2003.
19. Lech Polkowski and Piotr Artiemjew. On knowledge granulation and applications to classifier induction in the framework of rough mereology. *International Journal of Computational Intelligence Systems*, 2:315–331, 2009.
20. Philip Resnik. Using information to evaluate semantic similarity in a taxonomy. In Chris Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 448–452. Morgan Kaufmann, 1995.
21. Anjum Reyaz-Ahmed, Yan-Qing Zhang, and Robert W. Harrison. Granular decision tree and evolutionary neural svm for protein secondary structure prediction. *International Journal of Computational Intelligence Systems*, 2:343–352, 2009.
22. G. Salton. Developments in automatic text retrieval. *Science*, 253(5023):974–980, August 1991.
23. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, New York, 1983.
24. M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213. ACM, 1999.
25. Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124, 1998.
26. C. Shannon. A mathematical theory of communication. *Bell System Technology Journal*, 27:379–423, 1948.
27. Quan Thanh Tho, Siu Cheung Hui, Alvis Cheuk M. Fong, and Tru Hoang Cao. Automatic fuzzy ontology generation for semantic web. *IEEE Transactions on Knowledge and Data Engineering*, 18(6):842–856, 2006.
28. E. Voorhees and D. Harman. Overview of the Ninth Text REtrieval Conference (TREC-9). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the ninth Text REtrieval Conference (TREC-9)*, pages 1–14, Gaithersburg, Maryland, November 13–16 2000. NIST. Available from $http://trec.nist.gov/pubs/trec9/t9_proceedings.html$.
29. X. Yan, D. Song, and S. Li. Concept-based document readability in domain specific information retrieval. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 540–549, 2006.
30. J.T. Yao. Information granulation and granular relationships. In *Proceedings of the 2005 IEEE International Conference on Granular Computing*, pages 326–329, 2005.
31. Y.Y. Yao. Information retrieval support systems. In *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, pages 773–778, 2002.
32. Y.Y. Yao. Perspectives of granular computing. In *Proceedings of the 2005 IEEE International Conference on Granular Computing*, pages 326–329, 2005.
33. L. A. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8:338–353, 1965.
34. L. A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90:111–127, 1997.
35. X. Zhou, S.T. Wu, Y. Li, Y. Xu, R.Y.K. Lau, and P.D. Bruza. 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*, pages 558–564, 2006.