

Research on ontology-based literature retrieval model

Zhijun Zhang^{1, 2, 3}, Hong Liu^{1, 2*}

¹ School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China

² Shandong Provincial Key Laboratory for Novel Distributed Computer Software, Jinan 250014, China

³ School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China

Received 12 June 2014, www.tsi.lv

Abstract

Proper understanding of textual data requires the exploitation and integration of unstructured and heterogeneous scientific literature, which are fundamental aspects in literature retrieval research. The traditional literature retrieval is based on keyword matching, and the retrieval results often deviating from the users' needs. In this paper from the perspective of ontology, we built shareable and relatively perfect medical enzyme ontology, which is the foundation of the study of domain ontology constructing method. The ontology-based full text retrieval algorithm is put forward, and a document retrieval system based on medical enzyme semantics is designed and implemented, which can not only implement intelligent literature retrieval, but also improve the recall significantly while keeping high precision. This system can employ in particular area moreover it can be used in different areas of the semantic retrieval, which can provide intelligent foundation for the expert systems in medical enzymes field, information retrieval and natural language understanding, etc. The experimental results on the public medical enzyme domain dataset show that our approach performs better than the state-of-the-art methods.

Keywords: Ontology, Literature Retrieval, Semantic Web, Ontology Construction, similarity computation

1 Introduction

With the rapid development of computer network technology, the demand for information storage, transmission and processing power increases rapidly, and the retrieval and use of mass information has become an important research and application domain in computer information retrieval technology. Information retrieval is mainly implemented through search engines on the Internet, and its query function is to crawl on the Internet to retrieve resources by an automatic processing program using web spiders, and to visit public sites for resource collection and to organize and process the information correspondingly so as to provide users with convenient retrieval service.

Search engine is an indispensable tool for people to surf the Internet now, but with the wide use of search engines, users' satisfaction degree becomes increasingly low. Many results of information retrieval cannot meet the demands of users, which are either retrieval insufficiency or irrelevant. That is mainly because the current search engine generally adopts full-text retrieval technology based on keyword matching, which returns too much useless information and cannot reveal the semantic level of user queries. Semantic retrieval, which broke through the defects of mechanical matching confined to the surface, can understand and deal with users' retrieval request from the semantic level of words expression.

Tim Berners-lee put forward the concept of Semantic Web [1]. Its basic idea is that the data on the Web can be understood by machine through embedding the mark, which can be read by machine and represents some kind of knowledge in the creation and release of Web information. Semantic web [2] is considered to be the next generation of network technology, whose core is to use metadata to describe resources on the network, and it has been widely applied to knowledge retrieval in the area of library information. Knowledge retrieval emphasizes the semantic matching based on knowledge, while the ontology is just making a standard description and organization of knowledge from the semantic level with a good concept hierarchy. Full text retrieval algorithm based on ontology combined with the ontology knowledge logic can further improve the correlation between the information retrieval results and target, to make the retrieval results comply better with the needs of users [3].

With the improvement of medical scientific research level, when faced with more and more medical information resources, it is very difficult for people to understand them in time and apply them reasonably. In order to reduce the difficulty of finding effective information for the medical workers, as well as to improve work efficiency and avoid duplication of effort, we must use scientific methods to organize and manage medical information resources effectively [4]. The application of ontology is mainly used in the field of information retrieval and knowledge organization of

* *Corresponding author* e-mail: lhsdcn@jn-public.sd.cninfo.net

medical enzyme literature retrieval. Ontology can reflect restrictions between the mapping relationships of the lexical semantic and semantic which supports intelligent retrieval. The retrieval results will not be complete if the retrieval is done according to search terms provided by users only. When retrieving information, users usually hope that the results is something that they are interested in, and at the same time the engines could filter out irrelevant information, so that they could get the most valuable information. And when using ontology to retrieve information, researchers can use the ontology to map search terms to a set of specifications concept set automatically. Ontology can make a structured organization of information resources based on some knowledge organization system and can show links between content of information and knowledge organization system, and can connect domain knowledge base of ontology with information systems, so that in the process of using information, users can utilize ontology to understand specific concept and link the related resources more conveniently [5].

The remainder of this paper is organized as follows. In section 2, we provide an overview of ontology application in literature retrieval and related work. Section 3 introduces the construction of medical enzyme ontology and the overall framework of full-text retrieval system. Section 4 verifies the superior performance of full-text retrieval algorithm based on ontology in recall, precision and F-measure by the experimental data and results, followed by the conclusion and future work in section 5.

2 Related work

The traditional information retrieval is based on keyword matching, with the retrieval results often deviating from users' needs. At present semantic research in information retrieval mainly includes three aspects: natural language processing, method based on ontology and method based on concept. Voorhees first suggested using the concept of ontology to do query and expansion [6], and its basic idea is to use subclasses relationship of the ontology and synonyms. Ontoseek [7], developed by Guarino, is a retrieval system based on collaborative intelligent Agent. It can accurately describe the products or services in web pages, combining a content match mechanism driven by ontology with a formalization representation system with moderate expression ability, which tries to integrate ontology with big dictionary library, and provides users with a system in which interactive semantic query can be made with any words in the field. Although Ontoseek has quite effectively realized semantic functions, its degree of using ontology is not very high due to it is a content match mechanism. Swoogle [8] is a semantic web retrieval system based on the spider web concept, which extracts ontology from each searched text, comparing relationships between texts based on their ontology relevancy, but Swoogle method cannot search the

associated location ontology, as a result the ontology collection is not accurate. Maki puts forward the semantic retrieval methods based on the structure of ontology [9], effectively using the path in ontology to extend a user's query request. Navigli put forward a kind of query expansion method based on ontology annotation [10]. Literature [11] introduces an information retrieval model based on ontology: MELISA, which is used to retrieve literatures in the medical field. The Intelligent Information Processing Laboratory in the Institute of Computing Technology of Chinese Academy of Sciences established a kind of information retrieval server according to the theory of ontology and multiple intelligent agents [12], which can reflect dynamic changes of network information timely, and has good ability of information guidance.

At present, the discussion about the semantic web mainly concentrates on the research and development of the ontology. The concept of ontology initially originated in the field of philosophy, which was used to explain and illustrate the objective existence of things. With the continuous development of science and technology, ontology has been widely used in artificial intelligence, information retrieval, the semantic Web, natural language processing, and so on. In 1993 Gruber, who worked in Stanford University Knowledge System Laboratory (KSL), first presented a widely accepted definition of ontology: ontology is a specification of a conceptualization [13]. In 1998 Guarino proposed a concise definition of ontology and pointed out that ontology is a logical theory and is used to indicate a normal intended meaning of the vocabulary. Ontology is language-related; while concept is language-irrelevant. The concept of ontology has four layers of meaning: conceptualization, explicit, formal and share. Generally speaking, the ontology describes the relationships between the concepts in an application domain, which makes them to have uniquely definite meaning. With the help of ontology, we can obtain relevant knowledge of the field, and provide a shared understanding of the domain knowledge to facilitate communication between users and computers.

Ontology has become one of the effective tools for obtaining query expansion words. When using the method based on ontology for query expansion, we can select just a few extension words that are most closely related with the query expansion words by using the synonymous relationship, semantic entailment relationship, semantic extension relationship and semantic related relationship [14] between concepts. Association relationship between concepts is expressed by semantic similarity. By means of controlling the similarity threshold, we can adjust the scope of the extension concept set. Traditional models of semantic similarity calculation based on domain ontology between the concepts mainly has three types [15]: the semantic similarity calculation model based on semantic distance, the semantic similarity calculation model based on

content, the semantic similarity calculation model based on attribute. Domain-specific Ontology depicts both categories and instances in the field and their hierarchical relationships, inducts and abstracts the domain knowledge by defining elements such as categories, instances, attribute, relations, axioms and so on [16]. So far, many areas have emerged a large number of Ontology, such as medical ontology UMLS [17]. Ontologies that had been implemented mainly include: CYC, En-terprise, SENSUS, NKI, the massive knowledge system and medical knowledge library [18] presided by professor Cao Cungen, who works in the Institute of Computing Technology of the Chinese Academy of Sciences, and the research of software requirements elicitation method based on ontology developed by professor Jin Zhi at Beijing University and so on.

The above researches discussed the ontology retrieval model, but none of them involve ontology learning and inference problem, nor did they build their ontology models through formal ontology description language, which lead to neither fine nor precise use of ontology. At present there exists no large, shared, reusable, extensible medical enzyme ontology, thus the building of a good medical ontology is of vital significance. In this paper the medical enzyme ontology and its knowledge representation are built up based on the formal ontology theory. The ontology construction tools-Protégé is used to help building the ontology, which is the foundation through which knowledge acquisition and knowledge analysis are conducted based on medical enzyme ontology. Relevant function module has been realized in medical enzyme semantic literature retrieval system – MESLRS, which will provide intelligence foundation in the field of medical enzyme expert system, information retrieval and natural language processing.

3 The construction of ontology and framework design

3.1 ONTOLOGY

The ontology formalization description is as Equation (1):

$$O = C, R, H_c, Rel, A_o, \quad (1)$$

where C is the set of concepts. R is the set of relations. H_c shows the concept hierarchy, that is, the taxonomy relation between concepts. Rel shows the non-taxonomy relation. A_o is the ontology axiom. And here C and R is two disjoint sets. It can be seen from the structure of the ontology that the task of ontology learning includes the acquisition of concept, the acquisition of relations between concepts and the acquisition of axiom. These three kinds of ontology learning objects form the levels from simple to complex.

3.1.1 Ontology construction

Currently there is no unified method of the ontology construction and different methods are used in different research areas, so is the process of ontology construction. Now the way of ontology construction includes the following methods such as SEVEN, METHONTOLOGY, IDEFS, TOVE, FRAMEWORK, SENSUS, KACTUS and so on. At present many research fields have set up their own standard ontology, which indicates that the study of ontology model has entered a new stage. Generally speaking, the process of ontology construction can be divided into the following several stages: specification, conceptualization, integration, application and maintenance, in which knowledge representation, assessment and document management usually run throughout the entire process. Both knowledge representation and standardization are executed simultaneously in the first stage and evaluation is a stage of vital importance. In practice, ontology is usually constructed through method of accumulation, that is, a fundamental ontology is firstly constructed and then developed further. Many ontology constructions use a specific task as a starting point, which is easy for knowledge acquisition and descriptions of ontology function.

The early establishment and edition of ontology can only be conducted by experts and professionals of the field, while with the in-depth research and promotion of ontology, a series of ontology editing tools have been developed and each of which has specific advantages and disadvantages. Ontology development tools can be divided into six categories according to their different applications: ontology construction tools, ontology integration tools, ontology evaluation tools, ontology storage tools, ontology query tools and ontology annotations tools, and some of them might have multiple functions at the same time. Commonly used ontology development tools mainly include Protégé3.3.1, OntoEdit, OilEd, WebOnto.

3.1.2 Ontology learning

At present manual way of ontology construction is a kind of main ontology construction method, which is not only slow, but less efficient. In order to improve the efficiency of ontology construction, the concept of ontology learning is put forward.

Ontology learning [19] refers to the procedure during which the desired ontology is obtained from the existing data resource automatically or semi-automatically by means of machine learning and statistical techniques. It is still difficult to realize the completely automatic knowledge acquisition, and ontology learning can only be semi-automatically done under the guidance of users. Ontology learning includes the learning of concept, the learning of relationship and the learning of axiom. Ontology learning can be divided into the following three

categories: ontology learning based on structured data, ontology learning based on unstructured data and ontology learning based on semi-structured data. The current evaluation of ontology learning system has not yet formed a unified evaluation criterion.

3.2 ENZYME ONTOLOGY CONSTRUCTION

We take the industrial enzymes for instance to introduce the establishment of a medical enzyme. One purpose of establishing industrial enzymes ontology is to assist the teaching and scientific research, which can automatically show students the contents of knowledge to learn according to the process of the teaching, to help students to get more accurate query results and solve perplexed problems because it can provide a certain amount of semantic search on these resources. Therefore, we need to provide the semantic relations between concepts as much as possible. Industry enzymes ontology has not only stored the classification knowledge of the industry enzyme, the basic method of enzyme preparation but also stored the relationship between them.

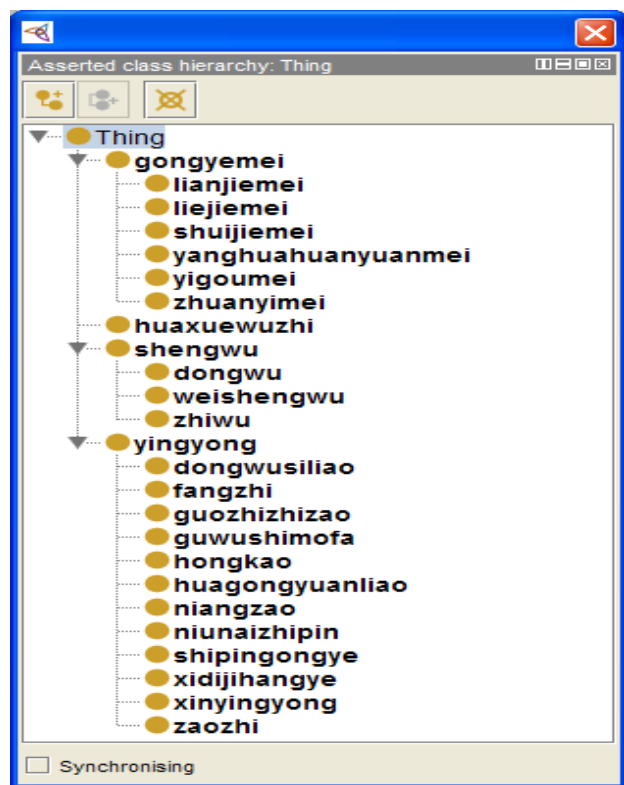


FIGURE 1 The graph of industrial enzyme hierarchical structure

The most important application of industrial enzymes ontology is the realization of data mining technology and the intelligent retrieval and mining of user requirement in literature systems. In this process, the validity and reusability of the industrial enzymes ontology model are also verified as well as the applicability of ontology as the knowledge organization system in the literature retrieval.

Zhang Zhijun, Liu Hong

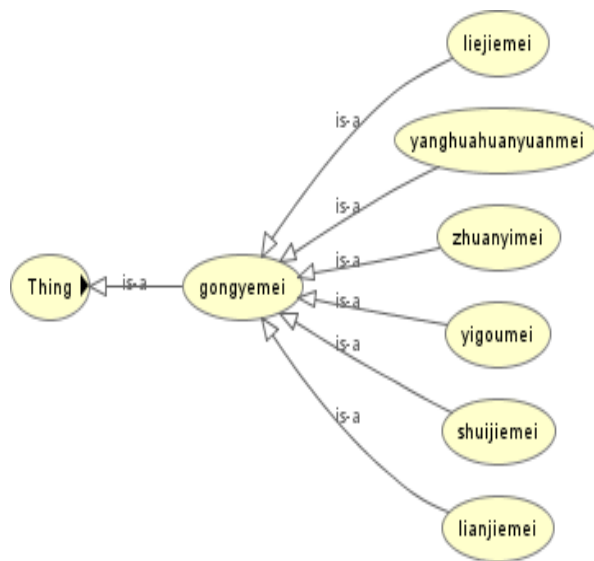


FIGURE 2 The graph of industrial enzyme classification tree

TABLE 1 OWL documents of industrial enzymes

```

<Ontology xmlns="http://www.w3.org/2006/12/owl2-xml#"
  xml:base="http://www.w3.org/2006/12/owl2-xml#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl2xml="http://www.w3.org/2006/12/owl2-
xml#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"
  xmlns:Ontology="http://www.semanticweb.org/ontologies/2013/1
1/Ontology.owl#"
  URI="http://www.semanticweb.org/ontologies/2013/11/Ontology.
owl">
  <SubClassOf>
    <Class URI="&Ontology;dongwu"/>
    <Class URI="&Ontology;shengwu"/>
  </SubClassOf>
  <Declaration>
    <Class URI="&Ontology;dongwu"/>
  </Declaration>
  <SubClassOf>
    <Class URI="&Ontology;dongwusiliao"/>
    <Class URI="&Ontology;yingyong"/>
  </SubClassOf>
  <Declaration>
    <Class URI="&Ontology;dongwusiliao"/>
  </Declaration>
  
```

With the help of plot plug-in graphviz-2.28 of Protégé4.0 a class relation diagram is automatically generated, as shown in Figure 1, 2, which shows the class hierarchy of industrial enzymes ontology. Owl: Thing is the superclass of all the classes. The four class of “yingyong”, “shengwu”, “huaxuewuzhi” and “gongyemei” are the subclass of the superclass-thing.

Part of the OWL documents described as Table 1.

3.3 FULL-TEXT RETRIEVAL ALGORITHM

3.3.1 Design of full-text retrieval algorithm

The query of users is often made up of a set of keywords, which are either elements of ontology library such as class, instance, attribute and attribute values, or other kind of common keywords. Therefore, the user’s input must be analysed firstly, which is done by using the analyser component in Figure 4. We use the same analysis algorithm for both indexing and retrieval, which can reach the optimal matching retrieval results. The design of full-text retrieval algorithm is as follows:

Algorithm 1: Full-text retrieval algorithm

Step 1: For each keyword in query, scanning the input keywords;

Step1.1: If the keyword is ontology element, then adding it to the ontology annotations stack h_2 ;

Step1.2: Else if keyword is other kind of keyword, then adding to the ordinary stack h_1 ;

Step 2: For h_2 to create space vector model v_2 , v_2 is given the higher weight value w_2 ;

Step 3: For h_1 to create space vector model v_1 , v_1 is given the lower weight value w_1 ;

Step 4: Integrating of v_1 and v_2 , creating user input space vector model v .

3.3.2 Basis of ranking algorithm design

Ranking of full-text retrieval algorithm is calculated based on the vector space model of information retrieval in Lucene system. For a collection of documents D , the closer between document d and query conditions q , the higher score of document d . The computation formula is as in Equation (2).

$$rank(q, d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t, \quad (2)$$

where $tf_{t,x} = \sqrt{termFrequency\ t, X}$,

$$idf_t = 1 + \log \frac{|D|}{documentFrequency(t, D)}$$

$$norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t^2}, \quad norm_d = \sqrt{|d|}$$

$$coord_{q,d} = \frac{|q \cap d|}{|q|}$$

3.3.3 Similarity measures method

In this section, we introduce some methods for similarity computation, with details of their adaptation to the ontology domain. We firstly exploit the geometrical model provided by concept hierarchies. The taxonomy

graph is shown in Figure 3. Wu and Palmer [20] (W&P) proposed a path-based method, which consider the distance between the concepts.

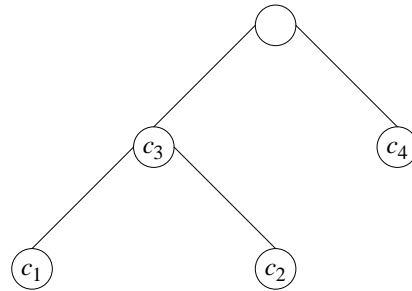


FIGURE 3 The graph of taxonomy

$$sim_{W\&P}(c_1, c_2) = \frac{2 \times N_3}{N_1 + N_2 + N_3}, \quad (3)$$

where N_1 is the number of is-a links from c_1 to its LCS (least common subsumer), N_2 is the number of is-a links from c_2 to its LCS, and N_3 is the number of is-a links from the root of the ontology to the LCS.

Li et al. [21] proposed a similarity method that compounds the depth of the ontology evaluated and the shortest path length in a non-linear fashion.

$$sim_{Li}(c_1, c_2) = e^{-\alpha path(c_1, c_2)} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \quad (4)$$

where $\alpha \geq 0$ and $\beta \geq 0$ are parameters, $path(c_1, c_2)$ is the shortest path length between concept c_1 and c_2 , and h is the minimum depth of the LCS in the hierarchy.

Choi and Kim [22] also proposed a similarity measure method, which is calculated according to the difference on the distance of the shortest path of two concepts and the depth levels between them.

$$sim_{CK}(c_1, c_2) = \frac{MAX_PATH - path(c_1, c_2)}{MAX_PATH} \times \frac{MAX_LEVEL - diff_level(c_1, c_2)}{MAX_LEVEL}$$

4 Experimental and data analysis

4.1 DATA SET

To test and verify the effectiveness of the full-text retrieval algorithm based on ontology, we developed medical enzyme semantic literature retrieval system – MESLRS. The purpose is to verify the correctness and reuse of the medical enzyme domain ontology model. MESLRS is constructed on .NET 2003 platform with SQL Server 2003 as the backstage database and deploying in the windows 2003 operating system. We selected 100 pieces of article about medical enzyme from cnki.net as text datasets of the domain, and used 12

categories of corpus in TanCorpV1.0 that is the corpus of Chinese text classification as background data sets.

4.2 METRICS

The test in this paper was based on the established medical enzyme ontology library. The evaluation metrics to evaluate the experiment result is *precision*, *recall* and *F*-measure, which are widely used in the field of information retrieval. The bigger of the value of the *precision*, *recall* and *F*-measure, the better of the result. Relevant concepts are shown in Table 2. Evaluation metrics are defined as in Equation (6) (7) (8).

$$Precision = \frac{A}{A+B}, \tag{6}$$

$$Recall = \frac{A}{A+C}, \tag{7}$$

$$F_j = \frac{2}{\frac{1}{R(j)} + \frac{1}{P(j)}}, \tag{8}$$

where R_j and P_j are the recall and precision of the document j respectively.

TABLE 2 Collection of documents measured by precision and recall

Full document collection		
	Retrieved documents	Not retrieved documents
Related documents	A (number of documents being retrieved correctly)	B (number of documents not being retrieve)
Unrelated documents	C (number of documents being retrieved wrongly)	Number of documents refused correctly

4.3 EXPERIMENTAL PROCEDURE

4.3.1 Full-text retrieval framework design

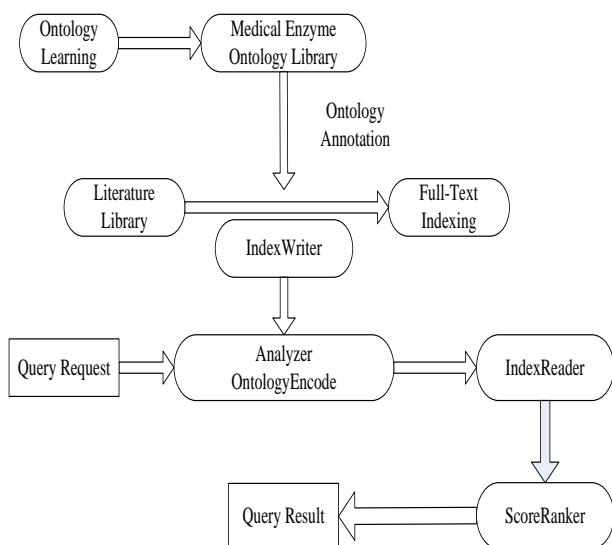


FIGURE 4 The retrieval model based on medical enzyme ontology

The full-text retrieval framework based on medical enzyme ontology model is divided into two parts: full-text index based on ontology and full-text retrieval based on ontology. Full-text index is the basis of full-text retrieval, because the indexed medical literatures are the objects of full-text retrieval. The reason for creating index is due to the vast data of the full-text, therefore in order to improve retrieval efficiency and integrate retrieval algorithm we must first create a full-text index and retrieval the full-text afterwards. In the full-text index mechanism based on ontology knowledge of domain ontology is integrated and indexed content is expanded by using the ontology annotations technology, which can s both recall and precision ratio of the retrieval. Similarly,

the full-text retrieval uses ontology library to analyse query request, which can grasp users' query request more accurately and provide them with a better retrieval ranking result. The full-text retrieval model is shown in Figure 4.

4.3.2 The implementation process of full-text retrieval algorithm

The implementation process of full-text retrieval framework base on medical enzyme ontology is as follows:

After reading a document from medical enzyme ontology library, the analyser component firstly executes annotations pre-treatment to content of the document, which will then deliver the annotated content the IndexWriter component to be indexed. The analyser component analyses the query request, matches its ontology elements with the ontology library, and returns the related literature. The ontologyEncoder component as a sub-component of the analyser component, switches the various elements in ontology library into a more efficient multi-way tree, which can not only be used for ontology annotations of IndexWriter component, but also be the scanning object when the IndexRead component queries the ontology elements. The ScoreRanker component is used to rank the literature of the query and list the ones that most conform to users' query request at the top to facilitate users to find the required documents quickly.

4.4 EXPERIMENTAL ANALYSIS

Although precision of traditional literature retrieval based on keyword matching is higher, it paid little attention to meanings and associations of words in different contexts for the reason that it did not consider the concepts that keywords might represent. However, the literature

retrieval algorithm based on ontology executes query expansions based on ontology, which compares the original query keywords submitted by users with the terms in the ontology library and finds the relevant words from ontology, then forms new query vectors so as to use

to full-text retrieval by adding the corresponding context relationship of keywords to the query keywords. So when evaluating the retrieval effect in this study we ought to consider the evaluation of semantic relevance.

TABLE 3 The comparison of experimental results

Semantic Relevancy	Recall (%)		Precision (%)		F-measure (%)	
	General	Ontology	General	Ontology	General	Ontology
0,1	2,73	4,79	99,33	100	0,05	0,09
0,2	4,9	8,87	98,12	95,15	0,09	0,16
0,3	12,47	19,92	99,56	95,14	0,22	0,33
0,4	16,63	35,05	98,01	94,99	0,28	0,51
0,5	21,98	47,36	91,83	94,03	0,35	0,63
0,6	30,59	69,93	89,38	93,33	0,46	0,8
0,7	36,65	81,85	84,95	88,2	0,51	0,85
0,8	53,18	90,25	71,84	76,01	0,61	0,83
0,9	64,18	98,96	58,67	56,93	0,61	0,72

In general query algorithm the recall and precision of the initial query term are calculated and we put the average value of several queries as the query results. Similarly in literature retrieval algorithm based on ontology the recall and precision ratio of the expand query terms are calculated and we put the average value of several queries as the results. Lastly, we calculate F-measure value of the general query algorithm and query algorithm based on ontology. The experimental results are shown in Table 3.

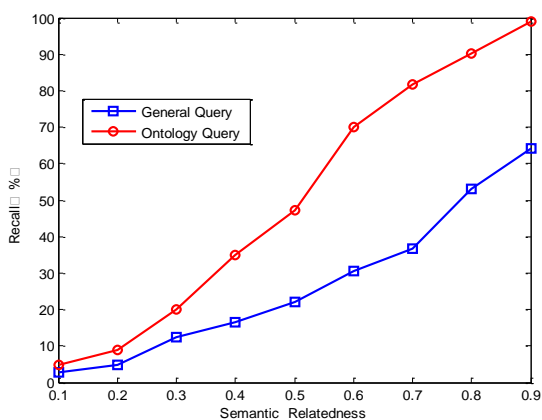


FIGURE 5 Recall comparison

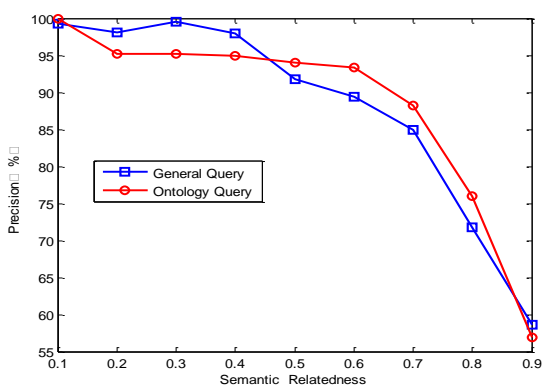


FIGURE 6 Precision comparison

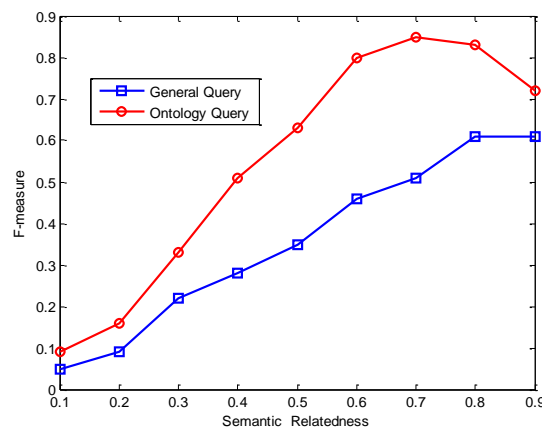


FIGURE 7 F-measure comparison

From Figure 5, 6, 7, we can draw the following conclusions.

Recall and precision has a reciprocal relationship in Figure 5, 6. When recall is high, precision is low. On the contrary, when precision is high, recall is low. A literature retrieval system can be compromised between them. In extreme cases, when retrieval system returns all documents of the system its recall is 100%, but the precision is very low. On the other hand, if the literature retrieval system can just return to the unique document, there will be a very low recall, but its precision could be 100%.

It can be seen from the Figure 6 that as far as precision is concerned the retrieval algorithm based on ontology has no special advantage. However, it can be seen from the Figure 5 that the recall based on ontology query is significantly higher than that of the general query. For example, when semantic relatedness is 0.6 the recall based on ontology query is 39% higher than that of the general query. Thus, the quality of query algorithm based on ontology is higher than general query based on keywords.

From Figure 7 we can find that F-measure of ontology query is higher than that of general query so it shows the advantages of ontology query.

Compared with the general query, the full-text retrieval algorithm based on ontology has the following advantages:

The full-text retrieval algorithm based on ontology can extract implicit knowledge. Using ontology knowledge library to extract potential keyword can retrieve keywords information that the user does not input, however the results might be very useful to users. From Figure 8 we can find that the number of retrieval document is less than 100 by checking with general retrieval containing keywords "gongyemei", but more than 900 documents can be retrieved by an ontology semantic extension.

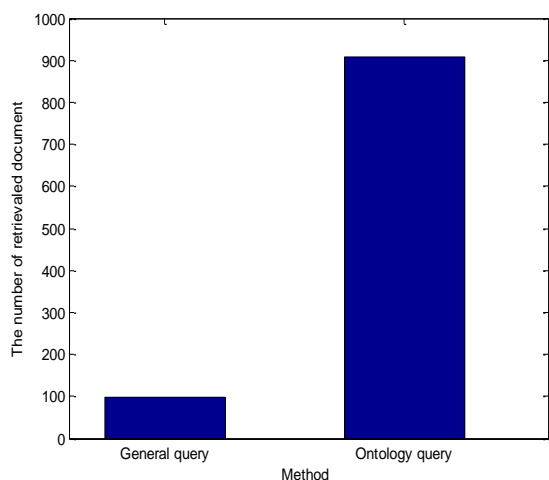


FIGURE 8 The number of retrieval document comparison with keywords "gongyemei"

The full-text retrieval algorithm based on ontology has intelligent query function. In the case of laccase we query the literature about functions of laccase. In medical enzyme ontology knowledge library, keywords and concepts are one-to-one correspondence, which guarantees that the query uses concepts instead of text. When the literature retrieval system based on ontology gets the keywords "qimei" and "gongneng" it firstly queries values of attribute "gongneng" of instances "qimei" in ontology knowledge library, with the returning values including the pulp bleaching, papermaking, hair dye, the effect of lignin, industrial wastewater treatment and so on. From Figure 9 we can find that the query result is 312 literatures on the application of laccase and the most appropriate literature is at the top by applying the ranking mechanism. On the contrary, the general query only find out less than 80 related literature when entering keyword "qimeidegongneng". There is no way of finding the literature corresponding to subordinate concept of "qimei" and "gongneng" no matter what kind of query retrieval strategies are used.

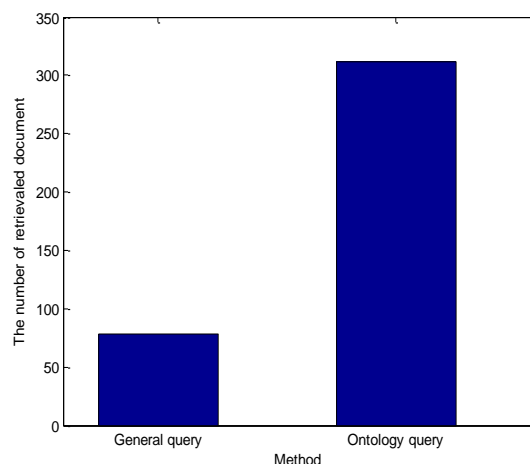


FIGURE 9 The number The number of retrieval document comparison with keywords "qimei"

5 Conclusions

This paper studies the construction method of domain ontology. We have constructed a reusable, shared, extensible, relatively perfect and practical medical enzyme ontology by using ontology to integrate and standardized medical enzyme knowledge system. The modification, query and storage of ontology library are carried out based on some applications of semantic web supported by Jena, and the full-text retrieval algorithm based on ontology is put forward. We have designed and implemented a literature retrieval system based on medical enzyme semantic, which have realized intelligent ontology learning about medical enzyme literature knowledge, as well as intelligent literature retrieval. In the premise of precision ratio, recall ration have significantly increased, what's more, F-measure of ontology query is higher than that of general query. It provides intelligent basis for medical enzyme literature related expert system, information retrieval, the education system, the natural language understanding and so on. This system is of strong reusability, because it cannot only focus on a particular area, but also be used in different semantic retrieval areas. As long as the corresponding domain ontology is changed the system can be used for information retrieval of the new domain. The efficiency of information retrieval is improved by using the advantage of ontology semantic expression.

In this article, we have just extracted part of the enzyme domain knowledge to construct the medical ontology model. Currently formalization of medical enzyme ontology is relatively rare and at the same time, there is no visual reasoning system, and the reasoning function of reasoning layer also needs to be strengthened. Therefore, it is very necessary for us to go on expanding and constructing enzyme ontology model, which is also our further research direction about ontology.



Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 61272094), Natural Science Foundation of Shandong Province (ZR2010QL01,

ZR2012GQ010), A Project of Shandong Province Higher Educational Science and Technology Program (J12LN31, J13LN11), Jinan Higher Educational Innovation Plan (201303001) and Shandong Provincial Key Laboratory Project.

References

- [1] Berners-Lee T, Hendler J, Lassila O 2001 The semantic web *Scientific american* **284**(5) 28-37
- [2] Spanos D E, Stavrou P, Mitrou N 2012 Bringing relational databases into the semantic web: A survey *Semantic Web* **3**(2) 169-209
- [3] Tsakonas G, Mitrelis A, Papachristopoulos L 2013 An exploration of the digital library evaluation literature based on an ontological representation *Journal of the American Society for Information Science and Technology* **64**(9) 1914-26
- [4] Jimenez-Castellanos A, Fernandez I, Perez-Rey D 2013 Biomedical Literature Retrieval Based on Patient Information *Biomedical Engineering Systems and Technologies* Springer Berlin Heidelberg 312-23
- [5] Lee C S, Jiang C C, Hsieh T C 2006 A genetic fuzzy agent using ontology model for meeting scheduling system *Information Sciences* **176**(9) 1131-55
- [6] Voorhees E M 1994 Query expansion using lexical-semantic relations *SIGIR '94* Springer London 61-9
- [7] Guarino N, Masolo C, Vetere G 1999 Ontoseek: Content-based access to the web *Intelligent Systems and Their Applications, IEEE* **14**(3) 70-80
- [8] Aleksovski Z, Klein M, Ten Kate W, et al 2006 Matching unstructured vocabularies using a background ontology *Managing Knowledge in a World of Networks*. Springer Berlin Heidelberg 182-97
- [9] Maki W, McKinley L, Thompson A 2004 Semantic distance norms computer from an electronic dictionary (wordnet) *Behaviour Research Methods, Instruments & Computers* **36** 421-31
- [10] Navigli R, Velardi P 2003 An analysis of ontology-based query expansion strategies *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia* 42-9
- [11] Abasolo J M, MELISA G M 2000 An ontology based agent for information retrieval in medicine *Proceedings of the First International Workshop on the Semantic Web Lisbon, Portugal* 73-82
- [12] Wu C G, Jiao W P, Tian Q J 2001 An information retrieval server based on ontology and multi-agent *Journal of Computer Research and Development* **38**(6) 641-7
- [13] Gruber T R 1993 A translation approach to portable ontology specifications *Knowledge acquisition* **5**(2) 199-220
- [14] Wang H, Sun R Z 2010 Research of semantic retrieval system based on domain-ontology and lucene *Journal of Computer Application* **30**(6) 1655-57
- [15] Resnik P 2011 Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language *arXiv preprint arXiv 1105.5444*
- [16] Zhong X Q, Fu H G, Yu L 2010 Geometry knowledge acquisition and representation on ontology *Chinese Journal of Computers* **33**(1) 167-74
- [17] Bodenreider O 2004 The unified medical language system (UMLS): integrating biomedical terminology *Nucleic acids research* **32** 267-70
- [18] Zhou X B, Cao C G 2003 Medical Knowledge Acquisition: An Ontology-Based Approach *Computer Science* **30**(10) 35-9
- [19] Du X Y, Li M, Wang S 2006 A survey on ontology learning research *Journal of Software* **17**(9) 1837-47
- [20] Batet M, Sánchez D, Valls A 2011 An ontology-based measure to compute semantic similarity in biomedicine *Journal of biomedical informatics* **44**(1) 118-25
- [21] Li Y, Bandar Z A, McLean D 2003 An approach for measuring semantic similarity between words using multiple information sources *Knowledge and Data Engineering, IEEE Transactions on* **15**(4) 871-82
- [22] Choi I, Kim M 2003 Topic distillation using hierarchy concept tree. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval ACM* 371-2

Authors	
	<p>Zhijun Zhang, born in 1973, in Shandong, China</p> <p>Current position, grades: Computer software and theory Specialty, Associate professor. University studies: He received his M. S. degree in School of Computer Science and Technology, Shandong University in 2006. Currently, he is a Ph.D. candidate in School of Information Science and Engineering, Shandong Normal University. Scientific interest: information retrieval and recommender systems. Publications: having issued more than twenty academic dissertations.</p>
	<p>Hong Liu, born in 1955, in Shandong, China</p> <p>Current position, grades: Professor. Doctoral supervisor. She is currently the dean of School of Information Science and Engineering, Shandong Normal University, Jinan, China. She is also the director of Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology. Publications: over 100 articles to professional journals. Scientific interests: Distributed Artificial Intelligence and Computer Aided Design.</p>