Novel Search and Retrieval Based on Domain Ontology

Hasnain Ali Salman, Xu De Zhi

GJCST Classification H.3.3

Abstract- **This paper proposes a framework of domain Ontology - based scientific and information retrieval (IR), and makes in - depth study on information organization and semantic retrieval. A prototype information retrieval information system is also implemented via a series of retrieval effects tests. The domain Ontology - based retrieval information system and the retrieval process of the information systematic prototypes are also depicted. In the end, the model is validated through a platform of trial information system. The result shows that when contrasted with traditional retrieval information system; this information system has a strong function in extending the connotation and denotation of the search words. So, it has the superiority in enhancing greatly the precision and recall.**

Keywords- Information retrieval (IR); Domain Ontology; Information organization; Semantic annotation; Semantic retrieval

I. Introduction

With the development of information technology, network-based environment of professional information resources increasing, and gradually evolved into a distributed, loosely network information environment. Information on the diversity of complex and heterogeneous information systems and so the resources of the professional literature, information organization and retrieval and use of proposed New challenges. Traditional information retrieval (IR) with the information system is mainly the professional literature to string matching and meta-data as a basis, but that expression match the same string Issues, rather than the meaning word form matching, vocabulary, and other isolated defects [1 - 4]; meta-data program is also the applicability of their capabilities and the evolution of different knowledge information systems and "Granularity"[2] of the resources there exist certain limitations described in, which have led to the current professional literature on information retrieval (IR) information system is less desirable[3]. As a new concept and method of information organization, ontology for the semantic retrieval theory provides an important way to solve the ideas and implementation. Since the ontology is cited in to the field of information retrieval since researchers in ontology construction [6], semantic annotation, retrieval, matching, similarity algorithms and search results in such areas as research, proposed Some distinctive ontology (IR) models and information systems [5 -7]. These studies indicate that contribute to the semantic level through ontology information retrieval, and to a certain extent, improve search results. However, current ontology-based information retrieval research focuses on enterprise information systems integration, Web Page retrieval and knowledge in the field, research priorities, but focused on search technology, reasoning logic and algorithm level, on the body of information retrieval in the professional literature Role, the information system architecture and implementation of less involved. As an important part of, (IR) professional literature has its own characteristics and research value [8]. Professional literature is not the same as unstructured text, thematic data Library contains a large number of staff by the use of indexing knowledge organization tools in understanding the original Based on the literature began to extract the contents of the semantic metadata[9-10-11]. In the information group Organization and information retrieval should take full advantage of the semantics of these meta-data value. And literature from the areas of information organization and heterogeneous concept, the concept query expansion and Semantic retrieval in various aspects of the implementation of the overall study. The paper is organized as follows: section 2 introduces related work on experimentation, while in section 3 we briefly introduce and the methods. The main contribution of this work is presented in section 4, which describes the experiments performed and the result analysis, and in section 5 that draws conclusions and outlines future work.

About- This paper is supported by the national natural science foundation of china (60970096)

II. Information Retrieval (IR) Model Based of Ontology

To solve the above problem, this paper specifically based on domain ontology Industry Information Retrieval model, its basic ideas are:-

1) With the current more mature information organizing tool to reflect Concepts and knowledge areas of the domain ontology.
2) In accordance with the domain ontology on the concept of the professional literature in the areas of progress Line automated semantic annotation, build the concept of the document vector.
3) When users search using the domain ontology of the query request extraction and semantic concept expansion of the concept of building a query vector.
4) Calculate the query concept vector and documents related to the concept of vector degree retrieve the relevant documents to meet client requirements.

The model framework Figure 1

Figure 1 Search model of professional literature based on domain ontology

Compared with existing research, retrieval model proposed in this paper mainly the following characteristics:-

1) In reference classification, topic and other traditional information organization method And meta-data methods based on the concept of ontology by constructing field, coding information System ontology and the formal description of the professional literature to the field of ontology concepts and heterogeneous integration and organization of the professional literature information.
2) The use of thesauri and classification to achieve the concept of ontology from the field Based on the dynamic transformation of semantic distance between concepts, semantic coincidence degree, Degree of relationship between the concepts of correlation is calculated to achieve a quantitative check Expansion.
3) The proposed relevance weights with semantic annotation method automatic processing without manual indexing of the professional literature, and expertise that have been indexing Industry documents, you can use its semantic metadata, semantic annotation to enhance effectiveness.
4) In the search stage, the proposed algorithm has certain adaptability, can be avoided because of incomplete ontology domain concepts influencing the cable results.

The following are the purposes of this organization in (IR) model, query processing, check Implementation in such areas as cable further elaboration.

1) *Professional Literature Information Organizations*

To achieve the professional literature of information organization, we takes the field was constructed concept ontology, ontology coding information system and the professional literature three inter-related body, domain and ontology. Among them, the concept of ontology to relevant areas of professional leader domain concepts and their hierarchical relationships, the use of natural words, etc For formal description coding information system ontology description of the professional literature Some external features of the above property values of the standardized, centralized management and maintenance of protection-related code, to ensure the encoding flexibility, scalability and reusability. Ontology defines the professional literature, professional literature class and is Resistance to the class instance to represent the professional literature and an example of real property are the external features of the literature, semantic information, and examples of inter-related documents formal description.

a) *Ontology*

The concept of ontology construction field is the basis of the whole information system and key High-quality domain ontology can bring better search results, but also means relatively high costs of ontology construction [1]. From the applicability, development is difficult degree of maturity and technical aspects, this article uses a thesaurus and classification to achieve the automatic transformation of ontology domain concepts [12-13-14], and also on the semantic distance between concepts, semantic coincidence degree, the concept of relevance, etc. Quantitative terms, the concept of similarity were constructed and the correlation matrix, a numerical Act sees Ontology coding information system is relatively simple, relevant standards and regulations [15] on the professional literature of the "type", "format", "Language", "time and space Range "and other attributes of the code are more clearly defined, by coding body Department of ontology on the definition of the formal description. Ontology construction in the professional literature, first established “Professional literature" Class as a top-level class, then according to relevant standards and norms [16-17] for the top layer type set up "Name", "Creator" "Association" And "Semantic Annotation" and other 16

properties, a number of attributes and then set the child property. For example, "Semantic Annotation". These attributes have "field concept" and "relevance weights". Two sub-attributes, "Professional Literature" was established under the top level category "Books" ,"Journal" ,"Conference paper", "Standard" , sub-class of 11 to represent a variety of documents types, these sub-classes inherit "Professional Literature" Class of all properties. Taking into account the special nature of various types of literature, but also the addition of these sub-categories Special attributes, such as "Conference paper" ,Class set", "Conference" Is a knowledge organization and knowledge management of, and contains The meeting name "TECHNOLOGY" ,"Meeting Date" ,,"Meeting place" and "Conference Name" and other sub-attributes; "Standard", Class from ,"Professional Literature" class inheritance to the "Other responsible person" Attributes increased "Approved body" Sub-attributes Ontology in the professional literature, each one corresponding to the professional literature, literature class model with an instance of the attribute value is the corresponding recording in the literature and Indexing items. "Language", "type", "format" and need to be standardized body from the coding information system of values to obtain, and the "field concept" property value is obtained from the field of ontology concepts. Use of "associated" attribute can also be real are examples of connections between different documents, such as a "journal paper" Real Cases can be "part" of a "journal" example; different versions of the book also can be "inherited version of" attribute to associate. Professional literature ontology can In order to better achieve different types of professional literature of information integration, and is very easy to expand the type and attributes.

b) *Semantic annotation of the professional literature*

Ontology building is completed; one must also mark out the field of ontology concepts each instance in the areas of literature and the relevance of the concept of weight, and save to the professional literature examples of "semantic annotation" property in the process shown in Figure 2 as follows:

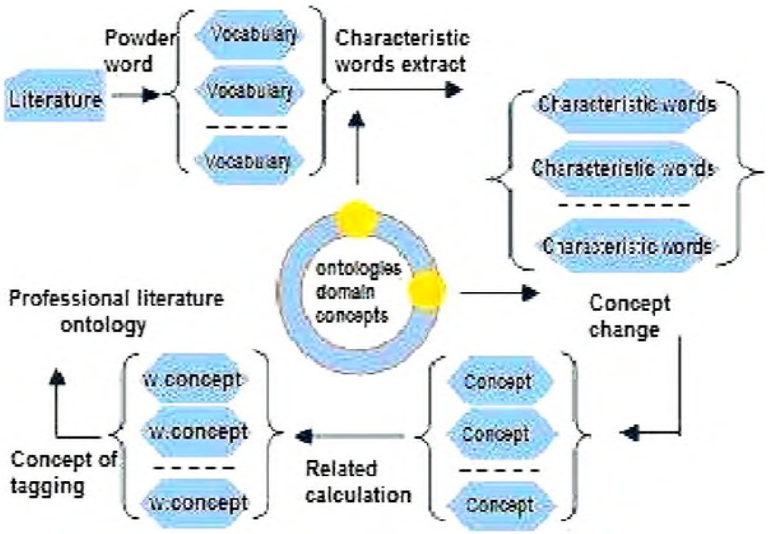


Figure 2 Semantic annotation process professional

Correlation weights in the calculation, considering the concept involved Words appear in the literature of different positions, in which documents and fields of long degree of several aspects to the concept of computational units, on the TF × IDF Operators Method [7, 18, 19] is improved, the concept of d c of the documents related to the right degree Value of w is defined as:

w = TF (c in d) · IDF (c) · field Boost (c. field in d) ·length Norm (C. field in d) (1)

Table 1 Correlation effects indicate that weight

Factors	Formula	Explain
Concept c in the literature d in the frequency of occurrence Rate	$tf(c \ \ in \ \ d) = \sqrt{frequency}$	Formula for the concept of frequency c in the literature in d frequency. Concept in the literature the frequency of a more High, then the concept of the higher correlation
Concept c in the whole Literature collection Reverse Frequency	$idf(c) = 10 + \ln \frac{numDocs}{docFreq + 1}$	num Docs the literature for the document collection Total doc Freq package for the literature collection Literature with the concept of the number of c. In a document collection, if the concept contains a The fewer the number of documents, Explain the concept of these The more relevant documents
Concept c appears in Literature in different fields The weighted value	Fields Boost(c field in d)= $\begin{cases} b1, C \text{ in Key word} \\ [b2, C \text{ in Title Keyword}] \\ b3, C \text{ in Abstracts} \\ b4, C \text{ in Other locations} \end{cases}$	The literature of a fixed set of terms b1, b2, b3, b4 are constants,respectively concept appears in the literature on different fields Correlation of factors. In general, Concept than appears in the title appears in the text Summary of important, out now surpasses Digest Important in the body of the
Concept c in which the word The length	Length Num(c field in d)= $\frac{1}{\sqrt{numfields}}$	Num Fields that c lies in the field The concept included in the total number of

Equation (1) the meaning of the various parts of the formula as shown in table i.

2) *Query processing*

Express the information needs of users, that search query is usually the form of the word. To realize the concept of matching based on semantic retrieval, on must deal with the conceptualization of the query request. Meanwhile, to increase the recall rate, also need to implement query expansion [20]. As the domain ontology contains a large number of concepts and their relations information, we use ontology to enter line a synonym expansion, semantic implication, the associated extension and semantic extension like [21]. In order to implement query expansion, first of all need to establish that the concept of inter- quantitative indicators of the degree associated the concept of similarity and relevance Matrix.

a) *The concept of similarity and the correlation matrix*

The concept of the concept of similarity that can replace compliance process level and semantic degree reflects the concept of their mutual aggregation characteristics. Conceptual similarity Is a very strong concept of subjectivity [22], from the perspective of different applications and start, the researchers made a variety of similarity calculation method [21], the considering the concept of culture between the structural level network diagram of a variety of factors, will The concept of ontology in the field of any two concepts X and Y define the similarity To:

$$Sim(x,y) = \frac{Size[P(x) \cap P(y)]}{Max\{Size(P(X)), Size(P(Y))\}} \bullet \frac{\alpha}{Dis(X,Y) + \alpha} \qquad 2$$

EQ (2) the right of the left part of the calculation of the equal sign is the language X and Y Justice degree, elements for the X and Y the number of concepts shared by the host, the denominator for the X or Y of the upper maximum number of concepts. Where, P(X) that takes study the concept of node X and all upper set, P(X) ∩ P(Y) table said the concept of the concept of X and Y by the concept of shared set of upper, Size (A) a number of elements within that set. Eq (2) the right of the right half of the equal sign calculate the semantic similarity of distance, where Dist (X, Y) for the X and Y of the semantic distance; α as regulatory factors, can approximate geographic Solution does not consider semantics in the case of coincidence degree, when the similarity is (0. 5) the semantic distance.

The correlation refers to the concept of linkages between the concepts together appears extent; reflect the characteristics of the combination between concepts. Taking into account the concept of and documentation related to the concept of degree of weight and is now in the literature of total frequency, almost Read X and Y can be defined as the correlation.

$$Col(X,Y) = \frac{\sum_{di \in D(X \cap Y)} W_{(X,di)} + W_{(Y,di)}}{\sum_{di \in D(X)} W_{(X,di)} + \sum_{di \in D(Y)} W_{(Y,di)}} \quad 3$$

Eq (3) WX,di said the concept of X in the document the correlation di Weight, WY, di , said the concept of Y in the document di correlation in weight. D(X) that all include the concept of X's document collection, D (X∩Y) that contains both X and Y are two concepts of the document collection, Eq(3) the concept of the molecular part of the said X and Y and all contain both of these two concepts the weight of the document and, to part denominator X and Y and for all with X or Y value of the document and the right. Definition of similarity between two concepts and the relevance algorithm, you can on the concept of ontology in the field of all the concepts two by two Calculate the similarity and phase Related degrees, and thus construct the concept of relevance and similarity matrix.

b) *Conversion and expansion of the semantic query*

Semantic query transformation and expansion of the query to retrieve the word form the concept of the request into the form of semantic information, and in accordance with the rules expand the concept. The process shown in Figure 3, the following steps, Method using word queries cut into several phrases, and with leading Extraction of domain ontology concepts related to the characteristics of the field of vocabulary.

According to the concept of ontology in the field of natural words the concept of identity, will feature words Meeting of the concept extracted, similarity threshold were set Rs (Rs∈[0, 1]) and correlation threshold Rc (Rc ∈[0, 1]), the control concept of similarity / correlation matrix, the domain concepts Ontology query request with the concept of similarity or correlation is greater than Rs greater than the concept of Rc extract, added to the original query, in order to achieve the requested query language Meaning extension. The query request cannot be converted into the concept of characteristic words are still words retain the form of exchange, to participate in subsequent retrieval.

III. Search Implementation

The implementation query request and quantification of the document collection and calculation. For the document to be retrieved set D in any text file dj ∈D, it can be expressed as s + t Dimensional vector of the form

$$d_j = (w_{1j}, w_{2j}, \cdots, w_{sj}, w'_{1j}, w'_{2j}, \cdots, w'_{tj}) \quad 4$$

Document vector by the concepts and vocabulary of two parts, Eq (4), s the number of concepts for the information system, t in the information system without a counterpart in the number of index terms. Wij for the concept ci of the document dj's relevance weights in on the text Offer examples for calculating the semantic annotation wij 'is the i-word ti on the document dj's relevance weights, using conventional TF × IDF method Obtained. For the expansion and transformation of the query request, it could be further form Show for the query vector.

$$q = (w_{1q}, w_{2q}, \cdots, w_{sq}, w'_{1q}, w'_{2q}, \cdots, w'_{tq}) \quad 5$$

Where s the number of concepts for the information system, t in the information system without the corresponding, the index number of the concept of the word. Vector component wiq i-a concept that ci In the query q In weight, and there wiq ≥ 0. wiq 'said the first i- Glossary ti in the query q in weight, and there wiq '≥ 0 Queries in the document and that the basis to quantify, document and The degree of similarity between the query can be transformed into two-dimensional to s + t Amount of similarity, and according to cosine law [19] to calculate the specifi Value.

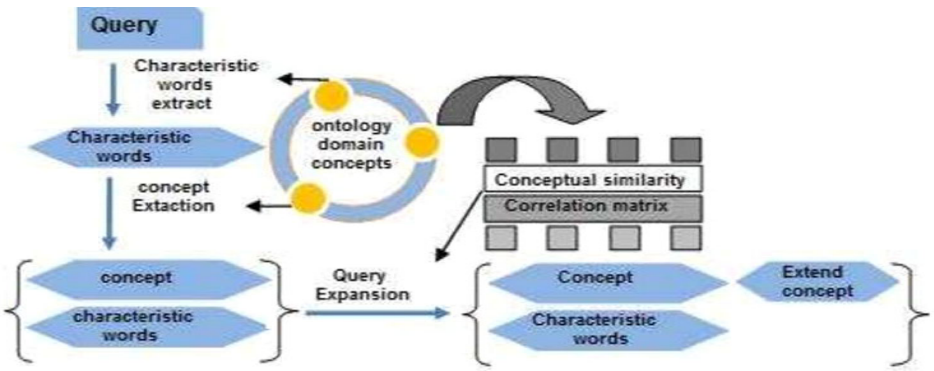


Figure 3 Semantic query transformation and growth process

IV. Evaluating IR systems

The methods are needed so as to be able to compare their abilities. In this paper you have two very different aspects of an IR system can be measured: efficiency, and effectiveness. Efficiency can be measured in terms of the resources required by the system, including the storage space required to store the document collection, and the computing resources needed to perform operations on the collection, such as the addition and removal of documents, and performing queries. Effectiveness attempts to measure, as the name implies, the effectiveness of an IR system at satisfying a set of queries. Given a sufficiently general document and query collection, the effectiveness should provide a domain neutral measure of the ability of the system. The measure of effectiveness is further complicated by the fact that it is dependent on the type of task being evaluated. Interactive systems must be evaluated in a different way to systems in which user feedback plays no role, Some of the issues have been discussed in the first international workshop on adaptive information retrieval (Joho et al.,2008) organized by the guest editors. They are Many IR researchers [Rijsbergen 2006] believe that a satisfactory approach to the evaluation of an information retrieval system is yet to be found. Since this is still a rich and ongoing area of research, we will only examine the most common evaluation methods. Furthermore, it is assumed that the system under evaluation operates with minimal user interaction. The most widespread method of evaluating an IR system involves providing precision-recall values for a set of queries posed on a specific document collection. Usually, a precision-recall diagram is plotted so that comparisons between IR systems can be made visually Use the document collection and a query, define R the number of relevant documents in the set for this query, and A be the number of documents retrieved by the IR system. Finally, let I be the number relevant documents within the documents retrieved by the IR system. Recall and precision can then be defined as

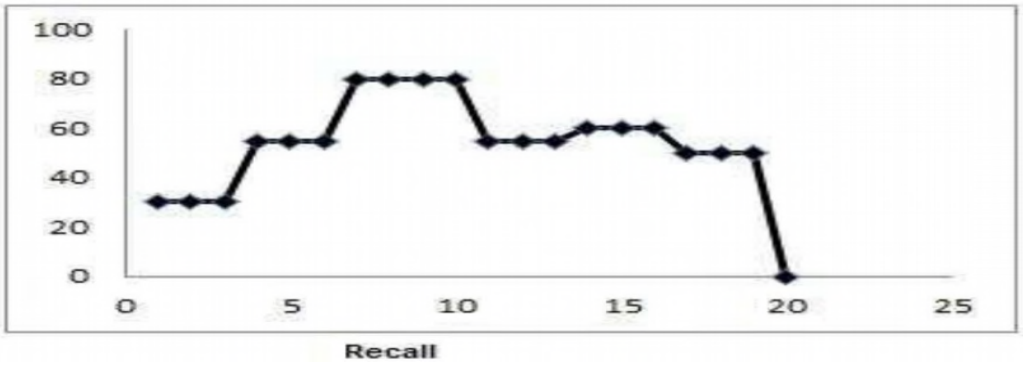
$$RECAL = \frac{I}{R}$$

$$PRECISION = \frac{I}{P}$$

The basic precision and recall measures assume that the IR system returns an unsorted list of results, which is then evaluated in full. If this is not the case, recall and precision values change as more documents within the result list are examined (the assumption is made that the list is ordered from the highest to least believed relevance). This is done by computing precision over the seen documents whenever a relevant document is found in the retrieved document list. For example, assume the following documents are retrieved in response to a query, with the documents marked by an asterisk indicating relevant documents:-

Rank	Document number	Relevant	Rank	Document number	Relevant
1	*dnu*19		11	*dnu* 77	*
2	*dnu* 40	*	12	*dnu* 53	
3	*dnu* 28		13	*dnu* 69	*
4	*dnu* 35		14	*dnu* 94	
5	*dnu* 3		15	*dnu* 100	*
6	*dnu* 9	*	16	*dnu* 65	
7	*dnu* 67		17	*dnu* 89	*
8	*dnu* 71	*	18	*dnu* 75	
9	*dnu* 43		19	*dnu* 13	*
10	*dnu* 36	*	20	*dnu* 5	

This paper is you assume that this query has 20 relevant documents, of which nine were retrieved as above. At a recall level of 6% (i.e.1 out of the 20 relevant documents have been seen), precision is 50%, since one out of the three seen documents are relevant. At the 60% recall level, precision increases to 50%. At 60% and 80% recall, precision values are 50% and 40%. Finally, at 90% recall, precision drops to 0%, as all relevant documents were not retrieved Precision-recall curves are normally drawn by computing precision at 20 standard recall values, namely, 0%, 5%, 10%... and 25%. If, as in the above example, insufficient relevant documents exist to compute recall at all these points, the values at the standard points are set as the maximum known precision at any known recall points between the current and next standard points. The example presented above would therefore yield the precision recall curve illustrated in Figure 4. Usually, precision-recall curves are computed by averaging the precisions obtained at the standard recall values over all queries posed to the system.



When averaged over a number of queries, precision-recall curves tend to follow an exponential decay curve. Intuitively this result from the fact that an algorithm would generally rank at least a few relevant documents quite highly, therefore yielding high precision for low recall values. As the number of relevant documents returned by the system increases however, more and more irrelevant documents are returned within the results. Obviously, a perfect algorithm would have a float precision-recall curve at the 100% level.

V. Conclusion

This paper analyzes the traditional information retrieval methods inadequate for achieving the semantic information system are described in four key modules, given A new ontology construction method based on this proposed ontology-based information system framework for information retrieval information systems, and describes the information system prototype Design and retrieval processes, effective solution to the traditional information retrieval recall and precision for low rate of problems; In addition, the paper also set up a A 100 or so computers in the field HTML format paper documents as a basis of knowledge of their search, by comparing the experimental results show that Ontology-based information retrieval compared to traditional information retrieval, it expanded the connotation and extension of search terms, making retrieval recall rate significantly Strengthened. Of course, the proposed ontology-based information retrieval information system prototype box leave many shortcomings, such as ontology and knowledge library of data increase, making information retrieval response time when the start search will be some effect, how to reduce the response time will be my they next step is to focus on the object of study.

VI. Acknowlegements

This work is supported by National Natural Science Foundation of China under grant No.60970096 and 90818004, the Postdoctoral Science Foundation of China under Grant No.20080440988 and the Natural Science Foundation of Hunan Province, China Under Grant No.09JJ403

VII. References

1) GÀbor NagypÀl. Imp roving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies[C]. In: Proceedings of OTM Confederated International Conferences, LNCS3762. Ber2 lin: Sp ringer - Verlag, 2005: 780 - 789.
2) Zhang, L., Yu, Y., & Yang, Y. (2005). An enhanced Model for Searching in Semantic Portals. WWW 2005, May 10-14, Chiba, Japan
3) .A.Varzi, Basic problems of mereotopology, in: N. Guarino (Ed.), Formal Ontology in Information Systems, IOS Press,Amsterdam, 1998
4) Paralic J, Kostial I. Ontology - based Information Retrieval[C] . In: Proceedings of the 14 th International Conference on Inform ation and Intelligent information System s, Croatia. 2003: 23 - 28.
5) Tomassen SL. Searchingwith Document Space Adap ted Ontologies [C]. In: Proceedings of Em erging Technologies and Inform ation information System s for the Know ledge Society, Athens, Greece. Berlin: Sp ringer - Verlag, 2008: 513 - 522.
6) Vallet D, FernándezM, Castells P. An Ontology - based Information RetrievalModel[C]. In: Proceedings of the Sem anticW eb: Re2 search and Applications. 2005: 455 - 470.
7) Motiee S, Nematzadeh A, ShamsfardM. A Hybrid OntologyBased App roach for Ranking Documents [C]. In: Proceedings of World Academ y of Science, Engineering and Technology, Prague. Czech Republic: World Acad Sci, Eng & Tech - waset, 2006: 55 - 60.
8) Li, Y., & Bontcheva, K. (2007). Hierarchical, perceptron-like learning for ontology-based information extraction. In Proceedings of the 16th international conference on World Wide Web (pp. 777-786). Banff, Alberta, Canada.
9) H. Ding, I. T. Solvberg,and Y. Lin, “A Vision on Semantic Retrieval in P2P Network”, in Proceedings of the 18th International Conference on Advanced Information

Networking and Application, 2004,pp.177-182
10) Kim, S.-B., Seo, H.-C., & Rim, H.-C. (2004). Information retrieval using word senses: Root sense tagging approach. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (pp. 258-265). Sheffield, UK
11) Hahn U. Turning Informal Thesauri into Formal Ontologies: A Feasibility Study on Biomedical Knowledge Re - use[J]. Com parative and Functional Genom ics, 2003, 4 (1) : 94 - 97.
12) Sean Bechhofer , Carole Goble, Thesaurus construction through knowledge representation, Data & Knowledge Engineering, v.37 n.1, p.25-45,April2001 [doi>10.1016/S0169-023X (00)00052-5]
13) Hepp M, De Bruijn J. GenTax: A Generic Methodology for Deriving OWL and RDF – S Ontologies from Hierarchical Classifications, Thesauri, and Inconsistent Taxonomies[C]. In: Proceedings of 1st International Conference onW eb Reasoning and Rule information System s,Innsbruck, Austria. 2007: 103 - 117.
14) Dublin CoreMetadata Element Set, Version 1. 1 [EB /OL]. [2008- 05 - 15]. http: / / dublincore. org/documents/dces/.
15) Li Peng, Tao Lan,"Measuring semantic similarity in ontology", Computer Engineering and Applications,vol.28, no.1,Jan.2007,pp227-229.
16) Lucene 2. 3. 1 Core AP I[EB /OL]. [2008 – 05 - 06]. http: / / lucene. apache. org/ java /2_3_1 / ap i/ core / index. html.
17) Bhogal J, Macfarlane A, Smith P. A Review of Ontology Based Query Expansion [J]. Inform ation Processing and M anagem ent,2007, 43 (4) : 866 - 886.
18) L. Dey, S. Singh, R. Rai and S. Gupta, Ontology aided query expansion for retrieving relevant texts, Proceedings of Advances in Web Intelligence, Third International Atlantic Web Intelligence Conference, Lodz, Poland (2005), pp. 126-132.
19) Couto, F. M., Silva, M. J., and Coutinho, P. M. (2007). Measuring semantic similarity between gene ontology terms. Data Knowl. Eng., 61(1), 137-152.
20) Formica A. Ontology - based Concept Similarity in Formal Concept Analysis [J]. Inform ation Sciences, 2006, 176 (18) : 2624 -2641.1987.
21) Zobel J. How Reliable are the Results of Large - scale Information Retrieval Experiments[C]. In: Proceedings of the 21st Annual International ACM S IGIR Conference on Research and Development in Information Retrieval. Australia: Melbourne. 1998: 307 - 314.
22) Zobel J. How Reliable are the Results of Large - scale Information Retrieval Experiments. In: Proceedings of the 21st Annual International ACM S IGIR Conference on Research and Development in Information Retrieval.. 1998: 307- 314.