Ontology based Similarity Measure in Document Ranking

Sridevi.U.K
Senior Lecturer
Department of Applied Sciences
Sri Krishna College of Engg and tech
Coimbatore, India

Nagaveni .N
Assistant Professor
Department of Mathematics
Coimbatore Institute of Technology
Coimbatore, India

ABSTRACT
This paper presents a methodology for the ontology based semantic annotation of web pages with annotation weighting scheme that takes advantage of the different relevance of structured document fields. The retrieval model is based on the importance factors of the structural elements, which are used to re-rank the documents retrieval by the ontology based distance measure. The relevance concept similarity are combined with the annotation-weighting scheme to improve the relevance measures. The proposed method has been evaluated on USGS Science directory collection. Preliminary experiments results show that our method may generate relevant document in the top rank.

Keywords
Ontology, Information Retrieval, Annotation, Semantic Search

1. INTRODUCTION
The rapid growth of documents, web pages and other types of textual content pose a great challenge to modern content management systems. Ontologies offer an efficient way to reduce the amount of information overload by encoding the structure of a specific domain and offering easier access to the information for the users. The degree to which different Web page elements are indicative of its content is in this paper referred to as significance indicator. The technique that determines the importance of different parts of a Web page improves the retrieval performance. However, all major ontology editors (such as Protégé[2], OntoStudio[3], are fully manual and offer little support to the users for structuring domains. Today's Web search technologies rely on link analysis techniques that exploit the structure of the Web to determine the important documents.
Because of the limitations of traditional retrieval mechanisms, conventional direct keyword based information retrieval technology cannot meet the growing user retrieval need with semantic knowledge. The keyword-based retrieval fails to integrate information spread over different resources. The information retrieval model does not utilize Semantics of the queries and document collection [7]. There have been many works which employ the Semantic Web technologies for information and retrieval such as TAP [4], KIM [11].A variety of aspects on improving search and ranking documents have been considered, such as concept based search of documents [9] [14]

The Semantic Web aims to achieve better data automation, reuse and interoperability [3]. The main advantage of Semantic Web is to enhance search mechanisms with the use of Ontology's [12].

Ontology is a general description of all concepts as well as their relationship. The Resource Description Framework /Schema (RDF(S)) and Web Ontology Language (OWL) are W3C recommended data representation models which are used to represent the ontology's [10] [18]. The basic method for constructing the Semantic Web is to use the terms defined in ontology as metadata to markup the Web's content. It is generally accepted that ontology refers to a formal specification of conceptualization.

In this paper, we propose an ontology based retrieval model for the exploitation of environmental sciences domain ontology's and knowledge bases, to support semantic search in document repositories. The research problem of improving relevance in search and ranking of documents requires techniques that consider the semantic annotation. The search system takes advantage of ontology based semantic annotation and it includes the weights of document structure in ranking [2]. Our current work is motivated by the need of new tools that can improve the retrieval and integration of information. In this context, we focus on ontologies whose specification components include entity classes, semantic relations among these classes, and distinguishing features that describe these classes. We evaluate our system on the collection of documents from USGS Science directory. Experimental results indicate that combining the ontology distance measure and semantic annotation weights improves the retrieval performance.

2. RELATED WORK
Over the last three years, the numbers of Semantic Web tools have been developed. The current research focus on providing a semantic metadata that enhances the information retrieval and support e-business applications. Automatic creation of metadata for Web pages resembles the task of semantic annotation in general[17]. A number of annotation tools for producing Semantic markups exist such as SHOE, Protégé, OntoAnnotate, MnM [4] [9]. There are several research projects about ontology-based information retrieval. The ontology definition of concepts can be used to describe the concepts and these concepts will be defined as document class [1]. SEAL [9] was conceived for semantic search of knowledge on the Web and has also been used for sharing knowledge on the WEB.

Semantic annotation is about assigning to the entities in the text links to their semantic description [10]. Annotation provides additional information about Web contents so that better decision on content can be made. Annotation ontology tells what kind of property and value types should be used in describing a resource. The usage of domain ontology's are employed for the annotations. To improve the recognition of important indexing terms, it is possible to weight the concepts of a document in different ways. For example, in topic indexing, concepts that form semantically related terms, gain more

weights. Although various annotation systems and methods have been developed, the question of how to easily and cost effectively produce quality metadata still remains largely unanswered. Dublin core annotation mainly describes properties of the document itself without providing too many details about its content. Ontology based annotations are instead developed to describe the content of the document and not its general properties. The manual annotation of document is a high cost and error prone task. To alleviate this task, an important effort is currently being made in automation of document annotation and the result is some degree of automation. However there is still some work to do achieve a complete automation of the annotation .The classical information retrieval model is incapable of supporting logical inference.

In general, most of the work about semantic annotation requires some predefined ontology's to extract, define and relate the annotation. The models of automatic semantic annotation are ontology driven semantic tagging and semantic meta data generation. The automatic semantic meta data generates Meta data that can semantically describe the content of annotating the page. The generated Meta data includes ontology by system defining its own semantic categories or a system relies on some predefined ontology. The annotation process of a web page is based annotating a web page with ontology and adding the relations between individuals. The ontology-based information retrieval based on vector space model describes the semantic annotation scheme of KIM platform [9]. It has reused automatic concept to label mapping available from the KIMKB [11]. In fully automatic annotation systems like KIM architectures support instance identification in a restricted predefined ontology model. Ontology based semantic annotation are needed when building the Semantic Web. The ontology-based information retrieval recognizes the relations among terms by referring to the ontology. Creating ontology's is not an easy task and obviously there is no unique correct ontology for any domain. The real quality of ontology can be assessed only for its use in real application. An ontology is a type of knowledge base that describes concepts through definitions that are sufficiently detailed to capturethe semantics of a domain. An ontology captures a certain view of the world, supports intentional queries regarding the content of a database, and reflects the relevance of data by providing a declarative description of semantic information independent of the data representation [11].

KIM [11] introduces a holistic architecture of Semantic annotation, indexing and retrieval for documents. It aims to achieve fully automatic annotation and to improve search and retrieval by integrating information extraction using GATE. Ontology based retrieval model work [9] complements KIM with a ranking algorithm specifically designed for ontology based retrieval model using semantic indexing scheme based on annotation weighting techniques. Genetic algorithms are generally quite effective for rapid global search to find solutions in nondeterministic problems. Genetic algorithm method enhances the efficiency and adaptability of a meta searching [15].Research in structural weights has suggested using document structures for document ranking. Genetic mining of HTML structures [15] uses the HTML tag weights to improve the performance of document retrieval system. The term that exists in title, bold, anchor tag adds more weights to the document then other terms. The document retrieval performance is improved depending on the structural importance of the document. Recent investigations in information retrieval and data integration have emphasized the use of ontologies and semantic similarity functions as a mechanism for comparing objects that can be retrieved or integrated across heterogeneous repositories[5],[6][8][16],[11],[14]. The distribution frequency of keyword improves the identification of important document based on the query [9]. The new similarity measure that incorporates the tag structure weights in addition to standard weighting schemes. Our approach combines the ontology similarity distance with annotation scheme to rank the annotated documents. The ontology based distance approach and annotation scheme significantly improves the retrieval performance especially for the top ranked document

3. ONTOLOGY BASED SIMILARITY

3.1 Document Representation

Document is composed of many terms and important words are spread out documents. The importance of significance indicators assigned to the Web elements like title, heading, bold, anchor improves the ranking of the Web documents. Unlike text documents, Web pages have certain characteristics such as structural information, hyperlinks and anchors which could serve as potential indicators of subject content. The relevance score of the document is assigned based on which term is matched and the part of the Web page in which the match is found. The annotation process of a web page is done with concepts of the ontology. Then, relations between individuals are discovered and instances are added. Document should be preprocessed to obtain semantic annotations and indexing of the document should be done. Examine the location of the annotated instance in the document. The annotation weights are calculated by combining the frequency and structures weights. Web search engines provide advanced features in that a user can specify how a query is matched with title of the page, text of the page, URL and links to the page, anywhere in the page. Although it is worthwhile to investigate how different matching affects the accuracy of the similarity of query to the document of retrieval results using tag weights [9]. Suppose a document collection on the Semantic Web is D= {d1, d2…,dn}. The number of occurrences of an instance in a document is primarily defined as the number of times the label of the instance appears in the document, if the document is annotated with instance and zero otherwise.

To extend vector space model to support structured ranking occurrences within each document structure must be included. The weight of a term in a document is basically computed by the classical tf.idf. The term frequency (TF) is the number of times that a term t appears in a document. The inverse document frequency (IDF) is the inverse of document frequency in the collection that contains term. The weighting scheme is defined in (1).

$$wk = tfk \cdot idfk \cdot \sum_{i=1}^{i=m} loc[tfk] \quad (1)$$

Where wk is the weight of kth term in the document, tfk is the frequency of the kth term in document, N is the total number of documents in the collection and idfk is the inverse document frequency annotated with kth term. loc [tfk] is the weight of the structural documents field.

3.2 ONTOLOGY DISTANCE MEASURE

The relations between entities are discovered through the measure of the similarity between the entities of ontology. The ontology based similarity between sets of concepts helps in retrieving and filtering information in automatic way.

In our paper, the methods are integrated to find the term relation information, while these terms are considered to be independent in the term-based vector space method. In order to find relation information between terms, first of all, we exploit the background knowledge which is given through an ontologies source WordNet. A matcher based on WordNet can be designed by translating the relations provided by the Wordnet to logical relations according to the rules of hyponym, hypernym, synonymy, antonymy relations. The terms s and t are related based on the ontology then the similarity between the terms is 1 otherwise 0. The simple measure based on synonymy similarity of Wordnet synsets is given in equation (2).

$$Sim(s,t) = \begin{cases} 1 & if\ s \cap t \neq 0 \\ 0 & otherwise \end{cases} \quad (2)$$

3.2.1 Resnik Semantic Similarity

Using Resnik semantic similarity [13] each synset of the concept is associated with a probability of occurrence of an instance of the concept in a particular concept. The lower probability is assigned to the more specific concepts. The Resnik semantic similarity considers the maximum information content and gives more general synset between the two terms. For Resnik the semantic similarity can be obtained per the frequency ofappearance in the corpus, and defined by :

$$Sim(x,y) = \max\left(IC(x,y)\right) \quad (3)$$

In (3) IC(Ci) = - log(p(Ci)) is the information content of the concept Ci (i.e, the entropy of a class Ci). The probability p(Ci) is computed by dividing the number of instances of Ci by the total number in the corpus. It provides, however, a systematic way to detect which entity classes are most similar to each other and, therefore, which entity classes are the best candidates for establishing integration across the ontologies. Our similarity measure could be used as a first step toward a strong integration of ontologies where user input would provide refinements. This approach is also useful in dynamic environments, such as the World Wide Web (WWW), where it may be impractical to force users to subscribe a priori to a shared ontology.

3.2.2 Minkowski Distance metrics

The similarity between the object can be analysed using Minkowski distance metrics. In (4) if p=2 then distances are the Euclidean distance. The distance measure weights give the importance of the dimension. Based on the importance of the concept the weights are assigned to the concepts and relations.

$$\forall x, y \in \delta(x,y) = \sqrt[p]{\sum_{i=1}^{n} \delta(x,y)^p} \quad (4)$$

.

3.2.3 Bayes Learning

The query term is matched with the concept in ontology using Bayes learning method. Based on Bayes rule in (5)

$$p\left(\frac{t}{c}\right) = \frac{p(c/t) \times p(t)}{p(c)} \quad (5)$$

Where t is the term in query or document and c is the concept in ontology. The probability of term annotated with the concept is considered in bayes rule.It provides, however, a systematic way to detect which entity classes are most similar to each other and, therefore, which entity classes are the best candidates for establishing integration across the ontologies. Our similarity measure could be used as a first step toward a strong integration of ontologies where user input would provide refinements. This approach is also useful in dynamic environments, such as the World Wide Web (WWW), where it may be impractical to force users to subscribe a priori to a shared ontology.

4. EXPERIMENTAL RESULTS

4.1 Test Collection

The keyword based analysis collects the set of keywords or terms that occur frequently together and then finds the correlation relationship among them. The semantic information retrieval KB has been built and associated to the information document base by using domain ontology's that describe the concepts. The query model can employ to find and manipulate the needful data from the annotated documents.The performance of the proposed methods are evaluated using web documents collected from USGS. According to USGS the topic major kind are environmental contamination, health and human impacts. The classification can be expressed by hierarchical structure of the class in the ontology definition. The concepts will be defined as the document class and some attributes, which describe the document information. The predefined base ontology described based on USGS Scientific directory provides the basis for the semantic

indexing of documents with nonembedded annotations. Documents are annotated with concept instances from the KB by creating instances of the annotation class. The semantic information retrieval KB has been built and associated to the information document base by using domain ontology's that describe the concepts. Once the experimental setting has been set up, we have tested the retrieval with IR functionality in GATE. GATE comes with a full-featured Information Retrieval (IR) subsystem. In Gate IR the documents can be retrieved from the corpora not only based on their textual content but also according to their features or annotations. The current implementation is based on the most popular open source full-text search engine – Lucene. The Ontology Annotation Tool (OAT) is a GATE plug-in available from the Ontology Tools plug-in set, which enables a user to manually annotate a text with respect to one or more ontology's. The required ontology must be selected from a pull-down list of available ontology's. OAT also allows users to assign property values as annotation features to the existing class and instance annotations.

4.3 Retrieval Performance measure

Our system takes the query and is executed against the knowledge base and returns the matching documents. A query weight gives the importance of the concept in the information needed by the user. The accuracy of the proposed technique has been evaluated against the result set generated by running the query " contamination water pollution"

Several measures such as precision and recall are used to evaluate the performance of document retrieval. Precision p is defined as the proportion of retrieved documents that are relevant and is given in (6) where Ra is the relevant document retrieved and A is the retrieved document.

$$P = \frac{Ra}{A} \tag{6}$$

Recall is defined as the proportion of relevant documents that are retrieved and is given in (7) where R is relevant document.

$$R = \frac{Ra}{R} \tag{7}$$

A is the number of retrieved documents. R is the number of relevant documents. Ra is the number of retrieved relevant document. A set of 20 queries was prepared manually for comparative performance measurement. The set of sample queries is given in Table 1. Table 2 show the different levels of performance for different cases the semantic information retrieval combined with the structural information improves the document ranking. Fig. 1 shows the performance of retrieval based on document annotation and without annotation for the query "minedrainage:Coal".

Table 1 Average Top-5 Search Ratings for 5 queries

Keyword queries	Keyword Rating	Ontology correlation rating
minedrainage:Coal	2.05	3.31
Contamination pollution:Water Quality	2.21	3.49
content:Acid Rain	1.25	3.35
environmental pollution	2.70	3.55
toxic	2.34	3.47

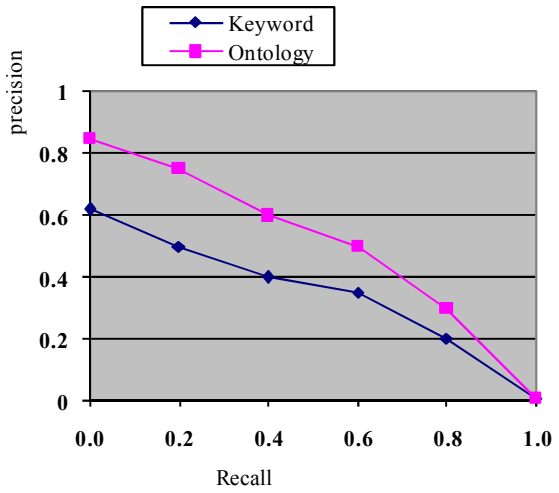

Figure 2. Comparision between keyword and ontology distance measure

The Table 2 shows the average weights for calculating the concepts weight. The Table 3 gives the precision and recall values. The performance of the two methods is compared using precision and recall. The F measure is defined using precision and recall shows the accuracy of the methods by comparing precision and recall.

Table 2 Average times for calculating concept weights

Number of Keywords	2	3	4
Time(msec)	6.31	9.40	12.50

Table 3. Best F measure using keyword and ontology distance measure

Methods	Precision	Recall	F-measure
Keyword	0.737	0.640	0.685
Ontology distance	0.768	0.727	0.747

Instead of simple keyword index lookup, the semantic search system processes a semantic query against the KB, which returns the relevant document. Better precision is achieved by using structured document annotation weight and the average precision for the top 10 documents is shown in Table 3. The Table 3 shows that the method can improve precision by 11% from 0.3761 to 0.4119 in relevant measure. Table concludes the ontology distance are more accurate compared to keyword.

5. CONCLUSION

As an extension of the current Web, Semantic Web provides a structured data and knowledge representation framework for Web information. Semantic Web provides a structured data and knowledge representation framework for Web information techniques to generate metadata that semantically annotating a web page will help improving lack of semantic information. These similar entity classes could be then analyzed with user inputs to derive semantic relations, such as is-a or synonym relations, to create a single, integrated ontology This paper introduces an annotation scheme that combines the ontology based similarity measures and concept frequency in the document Our approach can be seen as a evolution of the keyword based indices are replaced by ontology based KB and a semiautomatic document annotation weighting procedure that improves the retrieval performance.

6. ACKNOWLEDGMENT

The authors express their sincere thanks to the Management and Principal for their encouragement and support.

7. REFERENCES

1. Ahu SIeg, Bamshad Mobasher and Robin Burke, “Learning Ontology Based User Profiles: A Semantic Approach to Personalized Web Search”, IEEE Intelligent Informatics Bulletin, vol .8, pp.7- 18, 2007.
2. Dik L. Lee, Huei Chauang and Kent Seamons, ” Document Ranking and the Vector Space Model” , IEEE Software, pp.66-75, 1997.
3. Guan-yu LI, Sui-ming YU and Sha-sha DAI, “ Ontology based query system design and implementation”, International conference on network and parallel computing, pp.1010 -1015, 2007.
4. R.Guha, R.McCool and E.Miller, “Semantic Search”, International Conference on World Wide Web, pp.700-709, 2003.
5. N. Guarino, C. Masolo, and G. Verete, “OntoSeek: Content-Based Access to the Web,” IEEE Intelligent Systems, vol. 14, pp. 70-80,1999.
6. Jerome Euzenat, Pavel Shvaiko, "Ontology Matching", Springer-Verlag, Berlin Heidelberg(DE),2007,isbn:3-540-49611-4
7. Jose A.Alonso-Jimenez, Joaquin Borrego-Diaz, Antonia M. Chavez Gonzalez, Francisco and J. Martin-Mateos, “Foundational challenges in Automated Semantic Web Data and Ontology Cleaning”, IEEE Intelligent Systems, pp. 42-52, 2006.
8. J. Jiang and D. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” Proc. Int’l Conf.omputational Linguistics (ROCLING X), 1997.
9. J. Lee, M. Kim, and Y. Lee, “Information Retrieval Based on Conceptual Distance in IS-A Hierarchies,” J. Documentation, vol. 49, pp. 188-207, 1993.
10. Maedche, A, S.Staab, N.Stojanovic, R.Studer and Y.Sure, “Semantic portal: The SEAL Approach”, Spinning the Semantic Web, pp. 317-359, 2003.
11. Mehrnoush Shamsfard, Azadeh Nematzadeh and Sarah Motiee, “ ORank: An Ontology Based System for Ranking Documents”, International Journal of Computer Science, vol .1, pp.225- 231, 2006.
12. Pablo Castells, Mriam Fernandez and David Vallet,” An Adaption of the Vector Space Model for Ontology based Information Retrieval”, IEEE Transaction on Knowledge and Data Engineering, vol. 2, pp.261-22,2007.
13. Philipp Resink.” Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language”, Journal of Artificial Intelligence Research, Vol 11, pp. 95-130, 1999.
14. A. Smeaton and I. Quigley, “Experiment on Using Semantic Distance Between Words in Image Caption Retrieval,” Proc. 19th Int’l Conf. Research and Development in Information Retrieval SIGIR’96, 1996.
15. Sun Kim and Byoung-Tak Zhang, “Genetic Mining of HTML Structures for Effective Web Document Retrieval”, Applied Intelligence, vol.18, pp.243-256, 2003.
16. E. Voorhees, “Using WordNet for Text Retrieval,” WordNet: An Electronic Lexical Database, C. Fellbaum, ed., Cambridge, Mass.: The MIT Press, pp. 285-303. 1998.
17. Wang Wei, Payam M.Barjaghi and Andrzej Bargiela, “Semantic enhanced information search and retrieval”, Sixth International Conference on Advance Language and Web Information Technology, pp.218-223,2007.
18. Yufei Li, Yuan Wang, and Xiaotao Huang, “A Relation- Based Search Engine in Semantic Web”, IEEE Transaction on Knowledge and Data Engineering, vol.19, pp.273-282, 2007.