Effective Semantic Web Information Retrieval using Fuzzy Based Ontology

P. Nandhakumar, M. Hemalatha

Abstract— Information retrieval technology has been essential to the success of the Web. This paper presents a Fuzzy based Ontology approach which can improve semantic documents retrieval. Fuzzy Knowledge Base is being defined formally and new relationship called as semantic correlation is considered as a non-taxonomic relationship. This correlation is updated automatically when an insert is made into the database after processing a query. Here Information retrieval algorithm is used to find out the unique path between the entities being extracted from the queries by parsing. The extracted entities are used in construction of a SPARQL query using the template and the search is made to obtain maximum semantic association in the knowledge base. The implementation results obtained shows promising results in terms of precision and recall. The comparison made with other existing algorithms used has proved our proposed approach to be a outstanding one in information retrieval environment.

Index Terms—Semantic Web, Ontology, Fuzzy Theory, Information Retrieval.

I. INTRODUCTION

Nowadays information retrieval on web is based on keyword search which faces many drawbacks. The main drawback is that most web documents are in HTML, PDF,RTF formats which cannot be used for presentation, and hence the machine are not able to recognize the exact meaning of the documents being published [2]. Hence this drawback is overcome by using semantic web technique which makes use of RDF and OWL to describe a meaning for webpage information in a machine understandable format. The information is stored in a well defined manner, which enables the computers and users to extract the information easily. Recently many researchers have proposed increasing number of models based on keywords.

Ontologies are being constructed as models of reality to [1] satisfy some goals like

- Description and identification of relationship and objects in a specific domain that aids in share ability and reuse
- IR is made easier in this domain. The ontology is being constructed by the specialists of that particular domain and hence the end user may not understand the viewpoints of the specialists which creates problem and does meet the IR requirements.
- It is difficult to manage hundreds of concept in a single ontology

Manuscript received April, 2014.

P. Nandhakumar, He is working as Senior Software Engineer in Easy Design Systems Private Limited, Coimbatore, India.

M. Hemalatha, She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science at Karpagam University, Coimbatore, India.

Hence in this work Ontologies have been combined into a objects (XML database) which is used in searching of semantically correlated documents mentioned in the users query.

A. Introduction to fuzzy set theory and classification

Fuzzy sets are an extension of the classical sets and thereby have special membership levels. In classical set theory, the membership of elements in a set is assessed in binary terms according to a two-valued condition; an element either belongs or does not belong to the set [3]. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described by dint of a membership function valued in the real unit interval [0..1]. Therefore fuzzy sets generalize classical crisp sets, since the indicator functions of classical sets are special cases of the membership functions of fuzzy sets.

Fuzzy classification is an upgrading of traditional classification; equally fuzzy sets extend classical sets. The term classification describes the way of clumping elements into clusters, so that elements in the same cluster are as identical as possible, and elements in different clusters are as diverse as possible. In sharp classification, each element is associated with just one cluster. As a result the belonging of the elements to clusters are reciprocal and exclusive [5]. On the other hand fuzzy classification allows elements to belong to several clusters at the same time; and again like fuzzy sets, each element has a membership degree which reveals how far it belongs to the various clusters.

B. Fuzzy Theory

This section assesses some basic fundamental knowledge to be adopted in using this fuzzy theory [6].

DEFINITION 1 (FUZZY SET). A fuzzy set A on a domain U is defined by a membership function μ from U to [0.1], i.e., each item in A has a membership value given by μ. We denote Φ(S) as a fuzzy set generated from a traditional set of items S. Each item in S has a membership value in [0, 1]. S can also be called as a crisp set.

DEFINITION 2 (FUZZY RELATION). A fuzzy set A on a domain G×M, where G and M are two crisp sets is a fuzzy relation on G;M.

DEFINITION 3 (FUZZY SETS INTERSECTION). The intersection of fuzzy sets A and B, denoted as A∩B, is defined by μ A∩B(x)=min(μA(x), μB(x)).

DEFINITION 4(FUZZY SETS UNION). The intersection of fuzzy sets A and B, denoted as AUB, is defined by μ AUB(x)=max(μA(x), μB(x)).

DEFINITION 5 (FUZZY SET CARDINALITY). Let S be a fuzzy set on the domain U. The cardinality of S is defined as S , |S|=Σ μ(x),where μ(x) is the membership of x in S.

DEFINITION 6 (FUZZY SETS SIMILARITY). The similarity between two fuzzy sets A and B is defined as E (A, B) = |A∩B|/|AUB|

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. USING FUZZY CONCEPT IN INFORMATION RETRIEVAL

The semantic web is considered to be infrastructure that is being shared throughout and it consists of languages and tools for representing and processing the knowledge available in the web. The basic format used in the representation of knowledge is the Resource Description Framework (RDF) [2] and RDF SCHEMA [2]. The RDF data model helps in faster integration of data sources and bridges the semantic differences.

The RDF data model resembles similar to that of object oriented model which consists of entities that are represented using unique identifiers and also specifies the binary relationships or statements that can be developed between those entities.

In a graphical representation of an RDF statement, the source of the relationship is called the subject, the labeled arc is the predicate (also called property), and the relationship's destination is the object. Both statements and predicates are first-class objects, which means they can be used as the subjects or objects of other statements. The RDF data model distinguishes between resources, which are object identifiers represented by URIs, and literals, which are just strings. The subject and the predicate of a statement are always resources, while the object can be a resource or a literal [4].

The relationship between an object and an attribute is represented by membership value in [0, 1]. An ∞-cut can be set to eliminate relations that have low membership values. Generally, the attributes of a formal concept are considered as the description of the concept. Thus, the relationships between the object and the concept should be the intersection of the relationships between the objects and the attributes of the concept. Since each relationship between the object and an attribute is represented as a membership value in fuzzy formal context, the intersection of these membership values should be the minimum of these membership values [5-9].

Each fuzzy concept is associated with a membership function. There are many types of membership functions. Some of the common ones are [5]:

(1) Triangular. A triangular shaped curve can be described by three points, namely: (x1, 0), (x2, 1), and
(x3, 0). The RDF statements are as following:

```
<rdf: Description ID= "membership function 1">
<rdf: type resource= "# triangular"/>
<ex: points> <rdf: Seq>
<rdf: li resource= "# point1"/>
<rdf: li resource= "# point2"/>
<rdf: li resource= "# point3"/>
</rdf: Seq> </ex: points>
</rdf: Description>
```

(2) Trapezoidal. A trapezoidal shaped curve can be described by four points, namely: (x1, 0), (x2, 1), (x3,1), and (x4, 0). The RDF statements are as following[4]:

```
<rdf: Description ID= "membership function 2">
<rdf: type resource= "# trapezoidal"/>
<ex: points> <rdf: Seq>
<rdf: li resource= "# point1"/>
<rdf: li resource= "# point2"/>
<rdf: li resource= "# point3"/>
<rdf: li resource= "# point4"/>
</rdf: Seq> </ex: points>
</rdf: Description>
```

Using semantic relation defined in fuzzy linguistic variable ontologies between fuzzy concepts which can be the value of linguistic variables, fuzzy semantic information retrieval can be achieved.

In order to extend the query vector being automatically extracted to be used in information retrieval process a new algorithm based on fuzzy ontology has been proposed. When searching using this approach it is possible to find semantic links among the concepts: for each term specified in the query, a unique path is defined at each step, corresponding to the maximum value correlation.

III. PROPOSED METHODOLOGY

This work focuses on the ways to enhance the search process by making use of structured ontological data which represented by means of semantic mark up. The proposed approach works in several phases: Initially the user given query is parsed using the parser to analyze the query syntactically. Next the analyzed keywords extracted from the parser are send to the wordnet to find the relevant synsets. As a result semantically related words are obtained which are then used in automatic query generation process. This creates a refined SPARQL query which is used in the search process to obtain the relevant links. The extracted links are ranked to obtain the more precise links that are appropriate to the user query. The architecture of the proposed method is given in figure 1.

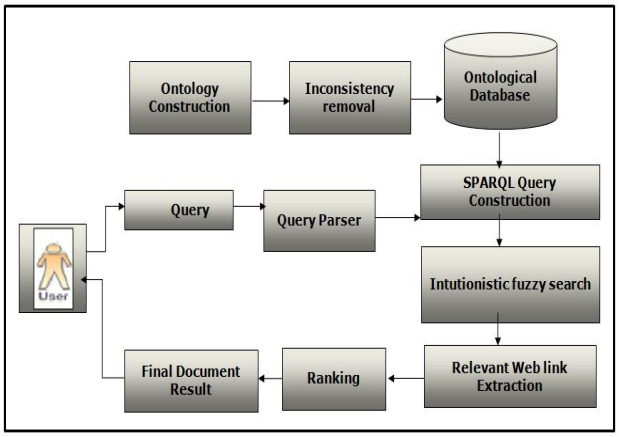

Figure 1: Proposed Architecture

A. Text Analysis

Query parsing: The input query given by the user is initially parsed by means of the parser. The parsing is done to analyze the query syntactically which helps to determine the part of speech of each and every word given in the query. By using this parsing technique the given query is analyzed grammatically.

Keyword analysis: The output obtained from the parser is sent to the wordnet to get the related synsets of various words contained in the query. By adopting this technique the semantically related words are obtained from the output of the wordnet. This Process is shown in figure 2 and 3.The figure 2 shows the user given query and figure 3 shows the parsing being performed on that user query.

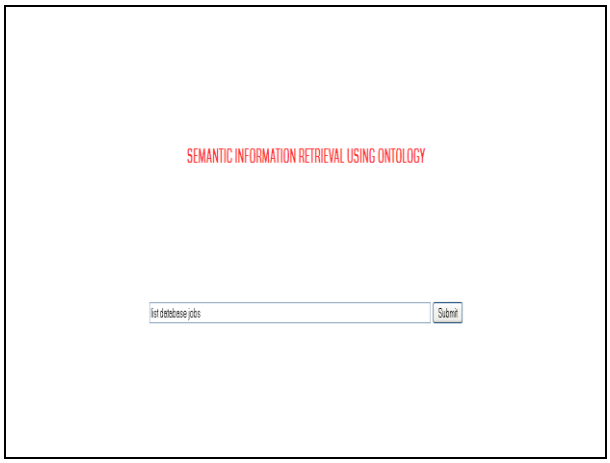

Figure2: User given Query

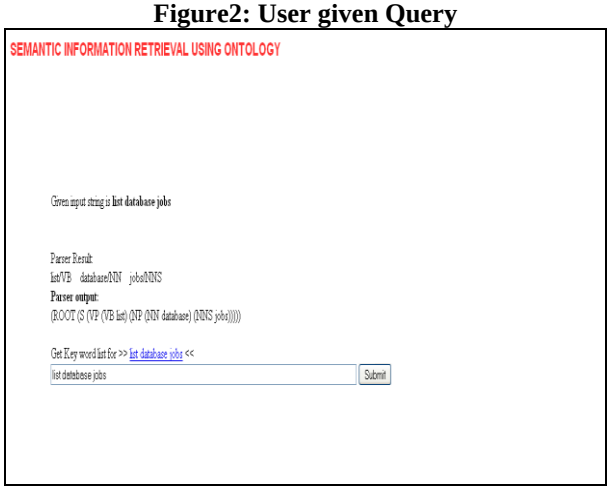

Figure 3: Text Analysis

B. Automatic Query Construction

The domain keywords that are semantically related to the query are extracted and the Refined SPARQL query is being constructed using the extracted keywords. The query is constructed automatically by using the predefined template which replaces the extracted keywords in the search domain. These refined queries are queries with expanded keywords and that has more semantic relevance involved. This process is depicted in figure 4. The queries formed will be more refined and will fetch more semantically related web links on passing these queries as input to the search algorithms.

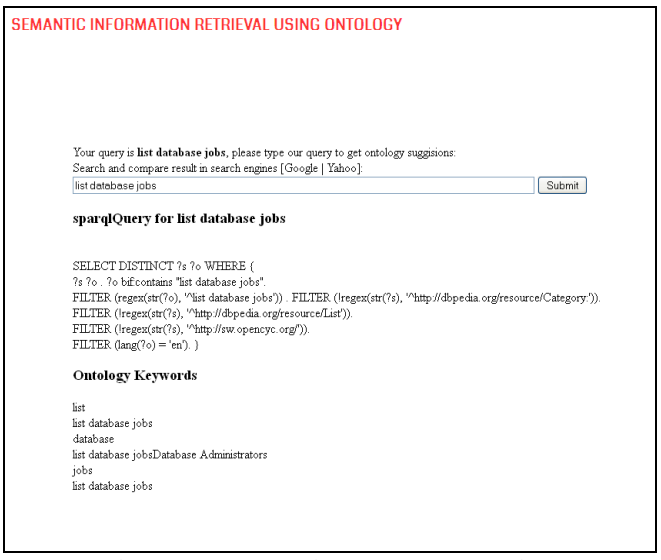

Figure 4: Automatic query construction

C. Query Execution and Process

The refined queries are sent to search fuzzy search Algorithm which fetches the web links related to the user query. The **fuzzy concept** has been involved in all the steps of the algorithm in order to semantically enrich the results that were obtained. The algorithm input is considered to be a input vector *Eq (i.e)* the terms being identified in the query.
Initially in this process the terms are used to locate the unique path finding maximum correlation value among them. *Eq* is extended navigating the **algorithm** recursively. Pruning is done directly to find immediately the important entities, which are more semantically correlated with reference to the *Eq* set.

D. Web Link Extraction

This process involves of more importance where the information related to the given user query is extracted from the domain knowledge base. The links extracted for a particular query is shown in figure 5.

E. Ranking

In this phase ranking concept is involved in order to directly extract the documents. Using the cosine distance among the weights of each entity the relevance score or the documents are calculated. This final score obtained is used to sort the links in the decreasing order. Hence using this score ranking among the documents is being performed.

IV. RESULT AND DISCUSSION

A. Evaluation Criteria

The proposed approach has been evaluated to know its performance. The performance of the proposed approach has been evaluated using Precision, Accuracy and Recall which is calculated as follows:

Accuracy

It is defined as the measure of how much of the information the system returns is correct is calculated as accuracy.

Precision Rate: Precision is the fraction of retrieved documents that are relevant to the search.

$$\text{Precision} = \frac{\#\text{ of relevant links given by the system}}{\text{Total }\#\text{ of links retrieved}}$$

Recall Rate: Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\left[\text{Re}\,call = \frac{\#\text{ of relevant links given by the system}}{\text{Total }\#\text{ of relevant links in web and proposed system}}\right]$$

B. Implementation Results

The proposed approach has been implemented successively in Xamp server and the results were collected by executing hundred queries simultaneously. The proposed method has also been compared with other previously existing algorithms like ANN and Genetic algorithm. The figure 5below shows the links being retrieved by executing a single query. The proposed approach has also been compared with common search engines like google and yahoo.

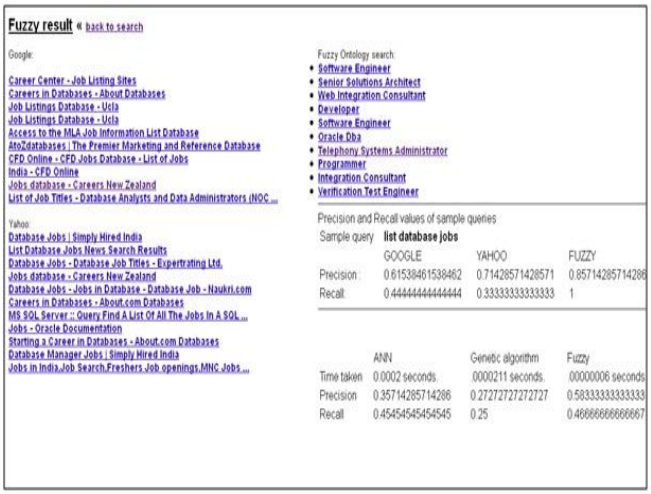

Figure 5: Implementation and comparison results

The result obtained in figure 5 shows that the number of links being extracted for the query is comparatively less than other common search engines and the precision and recall rate is also higher than other search engines. The Average precision and recall and total time taken is obtained by executing hundred query and the results obtained is shown in table1.

Table 1: Performance of the Proposed Approach

Method	Avg. Precision	Avg. Recall	Total Time taken	Accuracy
Proposed method	0.9875	0.6783	0.0016 sec	98.75
Common web	0.6136	0.2379	0.983 sec	61.36

The results obtained in table 1 clearly show that the average precision and recall rate for the proposed approach is higher than the value obtained for the common web. The accuracy of the proposed method shows promising values which shows the efficiency of the proposed approach.

C. Comparison Results

The proposed method has also been compared with the techniques that have been already used. Some important techniques like ANN and Genetic algorithm has been considered for evaluating the performance of the proposed approach. The performance is measured by executing hundred queries simultaneously. The result obtained is clearly shown in table 2.

Table 2 Performance Comparison Measures

Method	Avg. Precision	Avg. Recall	Total Time taken
Proposed method	0.9332	0.5648	0.0023 sec
Genetic algorithm	0.6986	0.4213	0.98380 sec
ANN	0.7435	0.5016	0.07235 sec

The result obtained clearly shows that the proposed method is much more efficient than other existing algorithms. The proposed method has higher precision and recall value and lower execution time.

V. CONCLUSION

The main goal of this work is to develop semantics-enabled Information Retrieval algorithm which can work better in terms of both syntax and semantics. A new conversion mechanism has been adopted to convert the query given in natural language to formal language after performing text analysis. The text analysis performed ensures that only semantically related terms are considered for query construction. The ranking mechanism has also been adopted to fetch the most relevant web pages for the query. The performance results obtained shows promising results in terms of accuracy, precision and recall. The obtained result has also been compared with other existing techniques to prove the efficiency of the proposed approach.

REFERENCES

1. N. Guarino and P. Giaretta, Ontologies and Knowledge Bases: Towards a Terminological Clarification.Toward Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 1995.
2. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific Am.,htttp://www.sciam.com/2001/0501issue/0501berners-lee.html, 2001.
3. Calegari, S., Farina, F.: Fuzzy Ontologies and Scale-free Networks Analysis. International Journal of Computer Science and Applications IV(II) (2007) 125–144
4. W3C, "Web Ontology Language Overview,"http://www.w3.org/ TR/owl-features/, 2006.
5. Sanchez, E.: Fuzzy Logic and the Semantic Web. Capturing Intelligence. Elsevier (2006)
6. Zadeh, L.: From Search Engines to Question-Answering Systems - The Problems of World Knowledge, Relevance, Deduction and Precisiation. In Sanchez, E., ed.: Fuzzy Logic and the Semantic Web. Capturing Intelligence. Elsevier (2006) 163–210
7. Calegari, S., Ciucci, D.: Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL. In: Proceedings of WILF 2007. Volume 4578 of LNCS. (2007).
8. Calegari, S., Ciucci, D.: Fuzzy Ontology and Fuzzy-OWL in the KAON Project. In: FUZZIEEE 2007. IEEE International Conference on Fuzzy Systems (2007).
9. Sanchez, E., Yamanoi, T.: Fuzzy ontologies for the semantic web. In Larsen, H.L., Pasi, G.,Arroyo, D.O., Andreasen, T., Christiansen, H., eds.: FQAS. LNCS 4027, Springer (2006) 691–699.

AUTHORS PROFILE

P. Nandhakumar, (nandhap@gmail.com) completed M.C.A., M.Phil. and currently pursuing Ph.D in computer science at Karpagam University, Coimbatore under the guidance of Dr.M.Hemalatha, Professor and Head, Dept. of Software System, Karpagam University, Coimbatore. He is working as Senior Software Engineer in Easy Design Systems Private Limited, Coimbatore.

Dr. M. Hemalatha, (csresearchhema@gamil.com) completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science at Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. She received best researcher award in the year 2012 from Karpagam University. Her research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing, Software Engineering, Bioinformatics and Neural Network. She is a reviewer in several National and International Journals.