Semantic Information Retrieval Model: Fuzzy Ontology Approach

Zeinab E. Attia
Computer Science Department, Institute of Statistical Studies and Research, Cairo University, Giza, Cairo

Ahmed M. Gadallah
Computer Science Department, Institute of Statistical Studies and Research, Cairo University, Giza, Cairo

Hesham A. Hefny
Computer Science Department, Institute of Statistical Studies and Research, Cairo University, Giza, Cairo

ABSTRACT
The paper proposes a multi-view information retrieval model. The model has the ability to deal with the multi-field topics problem using a predefined multi-field or multi-view fuzzy ontology. Respecting the natural relationship between concepts and terms, the model enhances the recall measure compared with previously proposed fuzzy ontology-based information retrieval models. It also proposes a ranking algorithm that ranks a set of relevant documents according to some criteria such as their relevance degree, confidence degree, and updating degree.

General Terms
Algorithms

Keywords
Information Retrieval, Fuzzy ontology-based information retrieval, fuzzy ontology

1. INTRODUCTION
An information retrieval system (IR) consists of a document collection, a user query, a retrieval engine, and a ranking module. It stores and annotates documents such that when users express their information needs in a query, the ranking module shows a set of ranked relevant documents. This set of documents is retrieved by the retrieval engine that associates a score to each one. The higher the score is, the greater the document relevance [7]. So, the challenge in IR is to find a set of most relevant documents respecting the user's query.

Researchers deal with this challenge using two different approaches. These approaches are keyword based approach and concept based approach. In the keyword based approach, documents are retrieved when they are annotated by terms specified in the searching query. However, this approach neglects many related documents that are not annotated with the query terms [7]. In the concept based approach, documents are retrieved according to their relevance to the searching query. This approach is a domain specific approach. It can be classified into ontology based approach and fuzzy ontology based approach. The performance of any IR system is measured using many computing parameters which are recall, precision, fmeasure… and many more [10].

Unfortunately, the current information retrieval systems suffer from many problems. Some of them are low in precision, low in recall and inability to deal with the multi-field topics problem.

Recall is the proportional of the correctly retrieved documents among the pertinent documents in the collection [11]. Precision is the proportion of the correctly retrieved documents among the documents retrieved by the system [11]. Multi-field topics are topics that combine two or more fields together such as the "bioinformatics" that combines the medical field with the computer science field. When certain medical user searches for a bioinformatics paper, the IR system will return the same set of documents that are returned to a computer science user. So these systems do not have the ability to distinguish between results of such topics respecting the field point of view.

The paper proposes a multi-view Fuzzy Ontology Information Retrieval model. It has the ability to deal with the multi-field topics problem. Also it aims to increase the recall measure respecting Leite [4] and FROM [5] models by considering the real relationship between concepts and terms in specific domain.

The rest of the paper is organized as follow; section 2 presents fuzzy ontology. Fuzzy ontology based Information Retrieval is discussed in section 3. Section 4 shows some related work. The proposed Linguistic based Fuzzy Ontology Information Retrieval model is presented in section 5. Section 6 shows a case study to test the proposed model. The paper is concluded in section 7.

2. FUZZY ONTOLOGY
Ontology is *"the conceptualization of a domain into a human understandable, machine readable format consisting of entities, attributes, relationships, and axioms"*. It is used as a standard knowledge representation for the semantic web [2]. Unfortunately, the conceptual formalism, supported by typical ontology, may not be sufficient to represent uncertain information commonly found in many application domains. This is due to the lack of clear-cut boundaries between concepts of the domains. Moreover, fuzzy knowledge plays an important role in many domains that face a huge amount of imprecise and vague knowledge and information, such as text mining, multimedia information system, medical informatics, machine learning, and human natural language processing. To handle uncertainty of information and knowledge, one possible solution is to incorporate fuzzy theory into ontology [8] yielding a fuzzy ontology model.

Accordingly, fuzzy ontologies contains fuzzy concepts and fuzzy memberships. Fuzzy ontologies are capable of dealing with fuzzy knowledge, and are efficient in text and multimedia object representation and retrieval [3]. There are many fuzzy ontology definitions according to the underlined application and domain. Some of them are:

[1] defines fuzzy ontology as a pair (C, R), where C is a set of concepts, R is a set of fuzzy relations between concepts.

[10; 11] defines fuzzy ontology as a quadruple(C, R, P, I), where C is a set of fuzzy concepts, R is a set of binary relations, P is a set of fuzzy properties of concepts, I is a set of individuals.

[14; 15] defines fuzzy ontology as a quadruple(C, R, F, U), where C is a set of concepts, R is a set of fuzzy abstract relations, F is a set of fuzzy concrete relations, U is the universe of discourse.

3. FUZZY ONTOLOGY-BASED INFORMATION RETRIEVAL

Fuzzy Ontology based Information Retrieval model, FOIR, is an IR model that semantically retrieves a set of relevant documents with respect to a certain query in a specific domain. This domain is represented using fuzzy ontology [5, 7, 8].Commonly, FOIR has three main components including input, retrieval processing, and output modules. The input module includes document collection, fuzzy ontology, and user's query. Retrieval engine and ranking module are retrieval processing components. The output component is the set of resulted ranked relevant documents. FOIR has four phases which are: document annotation, query expansion, retrieval of a set of relevant documents retrieval and ranking the set of resulted documents.

FOIR takes as input a set of documents, and a user query, to retrieve a set of the most relevant documents with respect to the entered query using a retrieval engine, then ranks this set and return it to the user. Both the document annotation process and the query expansion process depend on a fuzzy ontology.

4. RELATED WORK

Leite model [4] semantically retrieves a set of query's relevant documents in multi-domains. Each domain is represented as a fuzzy ontology and is then connected with other domains using fuzzy positive relations. It uses the well known "tfidf" method to annotate the document collection with a set of fuzzy ontology concepts. It deals with crisp queries. When a certain user enters a query, Leite expands it using a two phases query expansion process. The first phase expands each concept in the query with all of its related concepts in other domains. Then the result enters the second phase to expand each concept in it with all of its related concepts in the same domain. The max product composition between each document and the expanded user query is used as the similarity function to determine a set of the most relevant documents. This set of relevant documents is ranked in a descending order according to their relevance degree and returned to the user.

Fuzzy Relational Ontology Model, FROM, [5] is a document retrieval model based on fuzzy ontology. It semantically retrieves a set of relevant documents with respect to a user query. It assumes that each document in the document collection is already annotated with a set of weighted keywords. It considers fuzzy ontology as a set of concepts, terms, and relations between concepts and terms. FROM deals with crisp queries. When a user enters his query, it expands each concept in it with all terms that describes it and each term in it with all concepts that it describes. It retrieves a set of relevant documents using the max min composition between each document in the document collection and the expanded user query. The resulted set is ranked in a descending order according to each document relevance degree and then it returned to the user.

Fernández model [6] proposed an ontology based information retrieval model. This model deals with open environment. It annotates the document collection using two techniques. The first one is an NLP based, while the second is a context semantic information based. When a certain user enters a query, the model performs some processing on it using the ontology-based Question Answering (QA) system, PowerAqua. The adaptation of the traditional vector space IR model is used as to calculate the relevance degree of each document in the document collection with respect to the entered user query. Documents are returned to the user such that documents with higher relevance degree are listed first.

All of these models suffer from low in the recall measure, as a result of using incomplete fuzzy ontology components for expanding a certain user query keywords. Also, they cannot handle the multi-field topics problem. To rank the resulted documents, these models use the similarity degree between each document in the document collection and the user query keywords.

5. THE PROPOSED MULTI-VIEW FUZZY ONTOLOGY INFORMATION RETRIEVAL MODEL

The proposed model is a semantic document retrieval model that uses a predefined multi-view fuzzy ontology. It semantically retrieves a set of relevant documents according to a user's query respecting the underlined field or domain. It can be used to retrieve any kind of documents in a specific domain written in any language. The proposed model aims to:

- Increase the recall measure respecting FROM [5] and Leite [4] IR models. As its expansion algorithm uses a fuzzy ontology with components a set of concepts, relation between them, terms, relation between them, and a set of relations between concepts and terms.
- Deal with the multi-field topics problem. This is through using a predefined multi-view fuzzy ontology during its expansion algorithm to expand each user keyword in a certain field or view.
- Rank the resulted semantically relevant documents according to some criteria, such as the document matching degree, its confidence degree, and its timeliness.

5.1 The proposed Information Retrieval Structure

The proposed information retrieval model's main components are a set of annotated documents, users' profiles, users' queries, retrieval engine, and ranking module. It depends mainly on fuzzy ontology methodology and some NLP tools such as stemmer and POS tool.

Figure1 shows the structure of the proposed model. Firstly, users enter their query specifying the field search view. For example, select all papers about ***bioinformatics*** in ***computer science*** search point of view; here the user searches for papers about the keyword ***bioinformatics*** (keyword) according to the ***computer science*** search point of view. This query is then e expansion phase that expands each keyword with its related keywords using the predefined fuzzy ontology in its specified search point of view. Then, this expanded list enters the retrieval phase that semantically retrieves a set of matched documents each associated with a matching degree. This set is then ranked according to some criteria using the proposed ranking algorithm. Finally, the ranked relevant set of documents is displayed to the user.

5.2 The proposed Fuzzy Ontology Tool

The proposed fuzzy ontology model is a Multi-Views Fuzzy Related Ontologies, MVFRO [9]. Some of its main features are listed below:

- It is a general multi-domain fuzzy ontology, which can fit any domain and any application.
- The main fuzzy ontology components are concepts, relations, properties, terms, and individuals.
- In any domain, The relation between fuzzy ontology components or the related fuzzy ontologies can have multi-fuzzy-values each represents a certain point of view, e.g., In the old English, poetry represents the English literature with degree about 0.3, while in the modern English, poetry represents about 0.25 from the English literature.
- Using linguistic values and fuzzy number to express the relation between fuzzy ontology components or the relation between the related fuzzy ontologies.
- The used linguistic values and fuzzy numbers are defined by the domain expert according to his own subjective view.
- Storing all ontology components after stemming it in a relational database.
- Sorting different point-of-views that represent a certain relation between the ontology components or the related ontologies in one table instead of having one table per view.
- Storing the expert's subjective view about each used fuzzy number and linguistic term.

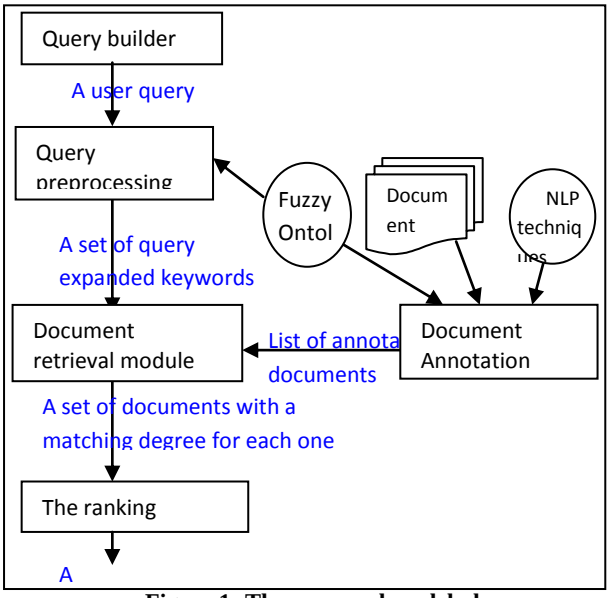

Figure 1: The proposed model phases

5.3 The proposed model phases

The proposed model phases are as follows:

1. Constructing a multi-view query

When a certain user enters his query, he should specify the underlined field and domain.

select all papers about bioinformatics according to the medical view

where "*bioinformatics*" is the keyword that the user searches for. "*medical*" is the search point of view. This linguistic term is previously defined by the user according to his subjective view and stored in his account.

2. Applying the Query Operations

After user submits his query, some operations are performed on it. First the query is parsed, such that each searched keyword is extracted with its field search view. Each keyword is then expanded in its specified search point of view using the predefined fuzzy ontology.

3. Retrieving a set of relevant documents

It semantically retrieves a set of relevant documents with respect to a certain user query through calculating document matching degree. A document matching degree is calculated as the max min composition between the list of weighted keywords that annotate this document and the list of query's weighted expanded keywords.

The result of this is a list of semantically relevant documents each associated with its matching degree.

4. Ranking the resulted documents

It ranks the resulted semantically relevant documents from the retrieval phase based on some criteria:

- The document's matching degree with user needs. The higher the matching degree is, the more document relevance with respect to user's needs.
- The document's confidence degree. This degree is extracted from the document's authors, the confidence degree of the journal, or conference that the document is published in. This factor reflects to what extent does the knowledge in this document is trusted. The higher the journal impact degree is, the more confidence that the knowledge in this document is correct,
- The document's updating degree. This degree is extracted from the document publishing date. This factor reflects to what extent does the knowledge in this document is new and updated, not out of date.

The ranked list of relevant documents is then displayed to the user in the same order.

6. APPLYING THE PROPOSED MODEL ON FROM CASE STUDY

This section applies the proposed model on FROM case study [5]. Figure 2 shows some changes in FROM fuzzy ontology. Considering the fuzzy ontology, it represents the computational intelligence domain in the theoretical point of view. Regarding fuzzy ontology structure, it also includes a set of relations between concepts and each other. All relations are represented as fuzzy numbers instead of membership degrees, for more realistic and accuracy in describing this relations. Consider the fuzzy number 'around' is defined by the expert using the triangular membership whose parameters 'a' and 'c' have the values '-0.1' and '+0.1' respectively. All the fuzzy ontology relations are interpreted using this definition then stored in the proposed model's database as in figure 3. Since the ontology size is small, the expert chooses inferring every new relation during its insertion time and

stores them into the database. This will decrease any ontology query response time.

Regarding FROM case study document collection, we assume each is annotated with a set of weighted keywords, a string of its authors, its published date, the conference or the journal that publishes it. Considering the set of weighted keywords, we will deal with the same set that FROM case study works on. For other annotations, we assume their values and store them into the document annotation database. Figure4 stores the document collection annotations in the database. For each document, we store its annotated weighted terms, weighted concepts, its publishing date, its authors, and journal or conference of publishing it.

Let's consider the following linguistic based query, Q:

Q: ***"Ontology" OR "Fuzzy Relation" in the theoritical view***

First, expand the user query as follow:

1-Check the first keyword type whether it is represented in the ontology as a term or a concept:

"***Ontology***" is a concept

2-expand the concept "**ontology**" in the theoretical of view as follow:

i- using its related concepts with degree >= 0.6

{(Information Retrieval, 0.8)}.

ii- using terms that describes it with degree >= 0.65

{(Taxonomy, 0.9), (Set Theory, 0.8)}.

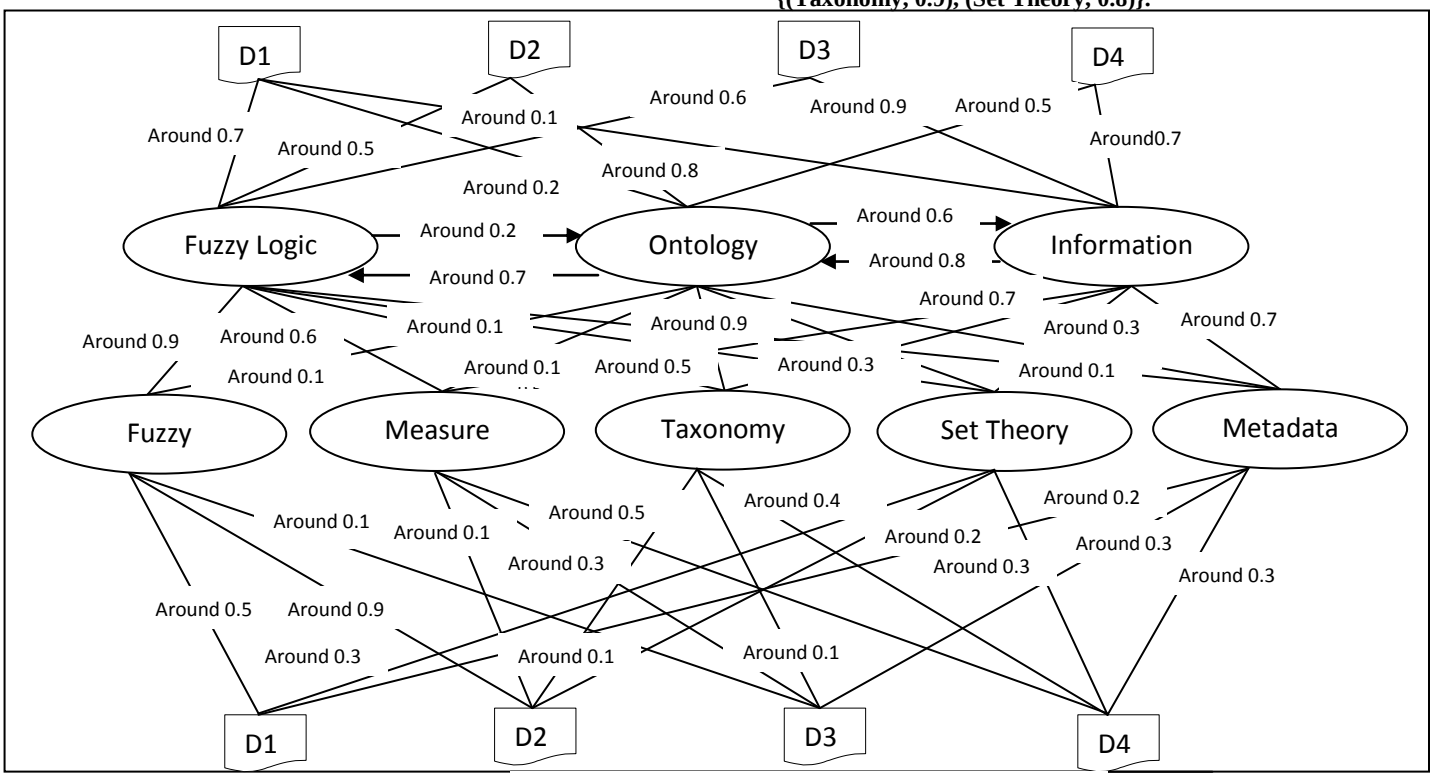

Fig. 2: applying the proposed fuzzy ontology on FORM fuzzy ontology [5]

Table 1: shows a list of journals each with its confidence degree

Journal name	Weight
International journal of intelligent systems	0.9
Knowledge and Information system	0.75
Advance in Fuzzy Systems	0.4

Table2: shows a list of authors and their confidence degree

Author name	Weight
M. A. A. Leite	0.8
J. Zhai	0.7
M. Hourali	0.3

iii- use the union operator between step 'i' and step 'ii' to have the expanded ontology set:

{(Information Retrieval, 0.8), {(set Theory, 0.8), (Taxonomy, 0.9)}.

iv- add the concept Ontology with degree 1 to step 'iii' to have the expanded ontology set:

{(Ontology, 1), (Information Retrieval, 0.8), {(set Theory, 0.8), (Taxonomy, 0.9)}.

3-Check the second keyword type whether it is represented in the fuzzy ontology as a term or a concept:

"Fuzzy Relation" is a term

4-expand the concept "Fuzzy Relation" in the theoritical point of view as follow:

i-using its related concepts that it describes with degree >= 0.65

{(Fuzzy Logic, 0.9)}.

ii-add the term fuzzy relation with degree 1 to step 'i' to have the expanded fuzzy relation set:

{(fuzzy relation, 1), (Fuzzy Logic, 0.9)}.

5-Apply the union operator on the expanded ontology set and the expanded fuzzy relation set:

{(Ontology, 1), (Fuzzy Logic, 0.9), (Information Retrieval, 0.8), (set theory, 0.8), (taxonomy, 0.9), (fuzzy relation, 1)}.

6-Divide the resulted expanded set into two sets, one for concepts and the other for terms:

Concept set= {(Information Retrieval, 0.8), (Ontology, 1), (Fuzzy Logic, 0.9)},

Term set= {(Set theory, 0.8), (Taxonomy, 0.9), (Fuzzy relation, 1)}.

Third, use retrieval engine to retrieve a set of relevant documents, each with its relevancy degree:

For each of the four documents,

7-Calculate the max min composition for the document concept set with the query concept set:

R_C= {(D1, 0.7), (D2, 0.3), (D3, 0.9), (D4, 0.7)}

8-Calculate the max min composition for the document term set with the query term set.

R_t = {(D1, 0.5), (D2, 0.9), (D3, 0.4), (D4, 0.3)}

9-Perform union operation on R_t and R_c:

R= {(D1, 0.7), (D2, 0.9), (D3, 0.9), (D4, 0.7)}

10-Apply the threshold on the resulted relevant document set with value 0.4:

R= {(D1, 0.7), (D2, 0.9), (D3, 0.9), (D4, 0.7)}

Class table

C_id	Class_name
11	Fuzzy logic
12	ontology
13	Information retrieval

Term table

T_id	T_name
1	Fuzzy relation
2	Measure
3	Taxonomy
4	Set theory
5	Metadata

Classterms table

T_id	C_id	Mship_deg	Is_infered	View_id
4	12	0.4	0	1
5	12	0.8	0	1
1	12	0.1	0	1
2	13	0.5	0	1
2	12	0.1	0	1
2	11	0.6	0	1
5	13	0.7	0	1
4	13	0.7	0	1
5	11	0.1	0	1
1	11	0.9	0	1
3	11	0.1	0	1
4	11	0.8	0	1
3	12	0.8	0	1
3	13	0.3	0	1

Classrelations table

C1_id	Rel	C2_id	Mship_deg	Is_infered	View_id
13	Related to	12	0.8	0	1
12	Related to	11	0.7	0	1
12	Related to	13	0.65	0	1
11	Related to	12	0.2	0	1

Fig. 3: Storing the fuzzy ontology in a relational database case study in database

Fourth, apply the proposed ranking algorithm:

For each document:

11-Calculate its confidence degree weight, assuming table 1 and table 2, using Eq. 1:

$$D_{Conf.Deg}=0.3*max(journal\_weight,author\_weight) \quad (1)$$

$D1_{Conf.\ Deg}$ = 0.27, $D2_{Conf.\ Deg}$ =0.12, $D3_{Conf.\ Deg}$ =0.21, $D4_{Conf.\ Deg}$ =0.12}

12-Calculate its updatence degree, using Eq. 2:

$$D_{update.Deg}=0.3*date\_weight \quad (2)$$

$D1_{update.\ Deg}$ = 0.27, $D2_{update.\ Deg}$ =0.12, $D3_{update.\ Deg}$ =0.21, $D4_{update.\ Deg}$ =0.12}

13-Calculate the document final weight, using Eq. 3:

$$D_{weight}=0.4*relevance\_degree+D_{Conf.Deg}+D_{updateDeg} \quad (3)$$

Relevance list= {D1= 0.82, D2 = 0.6, D3= 0.78, D4= 0.52}.

14-Rank the relevance list in a descending order as follow:

The resulted relevance document= (D1, D3, D2, D4)

As we can see, adding the relation between concepts and each other return document D2 as it is about fuzzy logic which is related to ontology.

Document table

Doc_id	Doc_name	Doc_path
1	D1	C://Documents//
2	D2	C://Documents//
3	D3	C://Documents//
4	D4	C://Documents//

Documentauthors table

Doc_id	Author_name
1	M. A. A. Leite
2	M. Hourali
3	J. Zhai
4	F.Hourali

Document publisher table

Doc_id	Publisher_name	Year
3	Advance in Fuzzy systems	2003
4	Advance in Fuzzy systems	2001
2	International journal of intelligent systems	2011
1	Advance in Fuzzy Systems	2002

Documentclassesannotation table

Doc_id	C_id	Mship_deg
1	11	0.7
1	12	0.2
1	13	0.1
2	11	0.5
2	12	0.8
3	11	0.6
3	13	0.9
4	12	0.3
4	13	0.7

Documenttermsannotation table

Doc_id	T_id	Mship_deg
1	1	0.5
1	4	0.3
1	5	0.2
2	1	0.9
2	2	0.1
2	3	0.1
2	4	0.2
3	1	0.1
3	2	0.3
3	3	0.4
3	5	0.8
4	2	0.5
4	3	0.2
4	4	0.7
4	5	0.3

Fig. 4: storing the annotation data for the document collection annotation

Table 3 shows a comparison between the proposed model and another two semantic-based IR models. The proposed model enhances the recall measure due to its query expansion algorithm respecting the underlined field and domain. It allows dealing with the multi-field topics problem through supporting the multi-field query tool.

Table3: shows a comparison between the proposed model and another two semantic TR models

The model	Handling the multi-field topics	Expansion algorithm		Ranking algorithm		
		Using synonyms	Using terms that describe it	Matching degree	Certainty degree	Timeliness degree
The proposed model	✓	✓	✓	✓	✓	✓
FROM [5]			✓	✓		
Leite model [4]		✓		✓		

7. CONCLUSION AND FUTURE WORK

This work presents an improvement in the fuzzy semantic information retrieval through:

- Retrieve a set of relevant documents semantically using the proposed fuzzy ontology tool MVFRO.
- Deal with the multi-field topics problem using a predefined mutli-view fuzz ontology.
- Rank the resulted set of documents according to some criteria which are their relevance degree with respect to use's query, confidence degree and updating degree.

The future direction to work in this area would be to build a document annotation algorithm using the proposed fuzzy ontology tool.

8. REFERENCES

[1] L. Dey, M. Abulaish, "fuzzy ontologies for handling uncertainties and inconsistencies in domain knowledge description," the 17th IEEE international conference on fuzzy systems, 2008.

[2] Q. T. Tho, S. C. Hui, A. C. M. Fong, T. H. Cao," Automatic Fuzzy Ontology Generation for Semantic Web," IEEE transaction on knowledge and data engineering, Vol. 18, No.6, June 2006.

[3] J. Zhai, Y. Liang, Y. Yu, J. Jiang, "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," JOURNAL OF SOFTWARE, VOL. 3, NO. 9, DECEMBER 2008.

[4] M. A. A. Leite, I. L. M. Ricarte, "Relating ontologies with a fuzzy information model," KnowlInfSyst, pp. 619-651, 2013.

[5] R. Pereira, I. Ricarte, F. Gomide, " Information Retrieval with FROM: The Fuzzy Relational Ontological Model," INTERNATIONAL JOURNAL OF INTELLEGENT SYSTEMS, VOL. 24, 340-356, 2009.

[6] M. Fernández, I. Cantador , V. López, D. Vallet, P. Castells, E. Motta ," Semantically enhanced Information Retrieval: An ontology-based approach," Web Semantics: Science, Services and Agents on the World Wide Web 9 ,pp. 432-452, 2011.

[7] M. A. A. Leite and I. L. M. Ricarte," A Framework for Information Retrieval Based on Fuzzy Relations and Multiple Ontologies," Springer, pp. 292-301, 2008.

[8] J. Zhai, M. Li, and J. Li, "Semantic Information Retrieval Based on RDF and Fuzzy Ontology for University Scientific Research Management," Affective Computing and Intelligent Interaction, AISC 137, pp. 661–668, 2012.

[9] Z. E. Alarab, A. M. Gadallah, H. A. Hefny, "An Enhanced Model For Linguistic-based fuzzy ontology," the 47th Annual Conference on Statistics, computer sciences and operation research, pp. 49-62, 2012.

[10] Y. Bassil, "A Survey on Information Retrieval, Text Categorization, and Web Crawling," Journal of Computer Science & Research (JCSCR) - ISSN 2227-328X , Vol. 1, No. 6, Pages. 1-11, December 2012.

[11] C. D. Manning, P. Raghavan, H. Schütze, "Introduction to Information retrieval: Evaluation in information retrieval," Cambridge University Press, 2008.