Fuzzy Ontology-Based Model for Information Retrieval

Zeinab E. Attia

Abstract—The paper proposes a linguistic-based information retrieval model. It has a linguistic-based query answering system to deal with a user linguistic-based queries in a certain field (view). Using these linguistic-based queries, users has the ability to define their needs accurately. The model also has the ability to deal with the multi-field topics problem using a predefined multi-field or multi-view fuzzy ontology. The model also enhances the recall measure respecting another two fuzzy ontology-based information retrieval models. The model also proposes a ranking algorithm that ranks the set of relevant documents according to some criteria such as their relevance degree, confidence degree, and updating degree.

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

I. 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 will show a set of ranked relevant documents. This set of documents is retrieved by the retrieval engine associating a score to each one. The higher the score is, the greater the document relevance [8]. So, the challenge in IR is to find a number of the most relevant documents according to 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 returned 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 [8]. In the concept based approach, documents are returned 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 [11].

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

Recall is the proportional of the correctly retrieved documents among the pertinent documents in the collection [12]. Precision is the proportion of the correctly retrieved documents among the documents retrieved by the system [12]. 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 linguistic-based Fuzzy Ontology Information Retrieval model. It has the ability to deal with the multi-field topics problem, allow its users to define their needs accurately. Also it aims to increase the recall measure respecting Leite [4] and FROM [5] models.

The rest of the paper is organized as follow; section II presents fuzzy ontology. Fuzzy ontology based Information Retrieval is discussed in section III. Section IV shows some related work. The proposed Linguistic based Fuzzy Ontology Information Retrieval model is presented in section V. Section VI shows a case study to test the proposed model. The paper is concluded in section VII.

II. 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
Accordingly, fuzzy ontologies will contain 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. Fuzzy ontology is a pair (C, R), where C is a set of concepts, R is a set of relations between concepts.

2. Fuzzy ontology is 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.

3. Fuzzy ontology is 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.

III. 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 components. The input component 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, retrieving a set of relevant documents, ranking a set of relevant 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.

IV. 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 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. Ranked Neuro Fuzzy Inference System, RNFIS, [7] proposed a hybrid information retrieval model. It is based on fuzzy version of vector space for information retrieval and fuzzy enhanced Boolean theory for document scoring. The model divides each document in the document collection into different weighted zones. When a certain user enters a query, the model expands it with multiple synonym queries based on its semantics. Then, it calculates the parameters; term frequency (tf), inverse term frequency (idf), and overlap (number of query terms found in the document) for each term of the expanded query in each zone in the document. These parameters in addition to the zone weight are fuzzified using the gaussian membership function. The retrieval engine uses these fuzzified parameters in Sugeno’s fuzzy rules. Then, it uses a certain aggregation operation to combine the result of all these rules. The result is then defuzzified to determine a crisp value. This result represents the relevance degree of this document respecting the user query.

All of these models can only deal with crisp queries. They are not able to deal with linguistic-based ones. So, its users can not define their needs accurately. Also, they suffer from low in the recall measure, as the 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.

V. THE PROPOSED LINGUISTIC BASED FUZZY ONTOLOGY INFORMATION RETRIEVAL MODEL

The proposed model is a linguistic-based semantic document retrieval model that uses a predefined multi-view fuzzy
ontology. It semantically retrieves relevant documents according to a user’s linguistic based query. It can be used to retrieve any kind of documents in a specific domain written in any language. The proposed model main features are:

- 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.
- Allow its users to define their needs accurately using linguistic-based queries, e.g., select all papers that are very related to bioinformatics in the computer science field search. “very related” is a linguistic term, “bioinformatics” is a keyword, “computer science” is the field or view point of search.
- Rank the resulted semantically relevant documents according to some criteria, such as the document matching degree, its confidence degree, and its timeliness.

A. 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, POS tool.

Figure 1 shows the structure of the proposed model. Firstly, each user should create a profile to define all his linguistic terms. Now, the user can build his query. This query is a set of keywords each is associated with its importance degree. This importance degree is expressed in linguistic terms. For example, select all papers that are very relevant to bioinformatics in computer science search point of view, here the user searches for papers that are very related (linguistic term) to the keyword bioinformatics (keyword) according to the computer science search point of view. This query is then passed on some operations, which are:

- Interpreting each linguistic term according to the user’s subjective view,
- Expanding 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.

B. The proposed Fuzzy Ontology Tool

The proposed fuzzy ontology model is a Multi-Views Fuzzy Related Ontologies, MVFRO [10]. 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.
- 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.

C. proposed model phases
The proposed model phases are as follows:
1. User profile creation
   User should create an account before building his/her query. Using this account, he/she can define any linguistic term according to his subjective view. Figure 2 shows the scheme of storing users’ linguistic terms definitions. When a certain user creates an account, this account is stored in the Users table. Then user can define any linguistic term, e.g., “related” is a linguistic term, using the userLinguisticTermFunction table. This table specifies which membership function is used to define a certain linguistic term according to the user subjective view. This membership function is also defined according to the user’s subjective view and stored in a table correspond to its name, e.g., triangularUserLinguisticTerms table for triangular membership function, piUserLinguisticTerms table for pi membership function. Hedges, e.g., very, more or less are hedges, can also be defined according to user’s subjective view and stored in userHedgeDefinition table. In userHedgeDefinition table user specifies the hedge name and its power. Also, users have the ability to specify which method is used to interpret a conjunctive or disjunctive query, by determining the conjunctive and disjunctive methods and storing them in userConjunctiveDisjunctiveMethod table. Now, user can build his query. Some query operations are then performed on this query.

2. Constructing a linguistic based multi-view query
   Now, the user can build his query. This query can be crisp, fuzzy, or linguistic based-query. For example, the query statement, “select all data mining papers” represents a crisp query. On the other hand, the query statement, “select all data mining papers with membership degree 0.6” and “select all data mining papers with membership degree around 0.6” are examples of the fuzzy query. A linguistic based query may be like:
   ```sql
   select all papers very related to bioinformatics according to the medical view
   ```
   where “bioinformatics” is the keyword that the user searches for. “very related” is a user linguistic term that reflects his needs for the keyword “bioinformatics”. “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.

3. 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 importance degree that is expressed using linguistic terms and hedges and with the search point of view. All linguistic terms and hedges are then interpreted according to the user’s subjective view. Each keyword is then expanded in its specified search point of view using the domain fuzzy ontology. The importance degree of any expanded word is the product of its relation with the original keyword and the importance degree of the original keyword.

4. 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 product 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.

5. 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.

VI. APPLYING OUR PROPOSED MODEL ON FROM CASE STUDY
This section tests the proposed model on FROM case study [5]. Figure 3 shows some changes in FROM fuzzy ontology. Considering 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
The proposed model’s database as in figure 4. Since the ontology size is small, we choose inferring any new relation during its insertion time and store 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. Figure 5 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 very related” OR “Fuzzy Relation related in the theoretical point of view”

First, interpret each linguistic term in Q using the user definition:

\[
\text{Related} = 0.88, \quad \text{Very related} = 0.79
\]

So, user query can be rewritten as

“Ontology 0.79” OR “Fuzzy Relation 0.88 in the theoretical view”

Second, 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 multiplied by 0.79 >= 0.6

\{(Information Retrieval, 0.632)\}

- ii- using terms that describes it with degree multiplied by 0.79 >= 0.65

\{(Taxonomy, 0.711), (Set Theory, 0.632)\}

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

“Fuzzy Relation” is a term

4-expand the concept “Fuzzy Relation” in the theoretical point of view as follow:

- i-using its related concepts that it describes with degree multiplied by 0.88 >= 0.65

\{(Fuzzy Logic, 0.792)\}

- ii-add the term fuzzy relation with degree 0.88 to step ‘i’ to have the expanded fuzzy relation set:

\{(fuzzy relation, 0.88), (Fuzzy Logic, 0.792)\}

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

\{(Ontology, 0.79), (Fuzzy Logic, 0.792), (Information Retrieval, 0.632), (Set Theory, 0.632), (Taxonomy, 0.711), (Fuzzy Relation, 0.88)\}

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

VII. CONCLUSION AND FUTURE WORK

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

- Building multi-views Linguistic based query system. This gives users more flexibility while building their queries.
- Allow users to define all their linguistic terms according to their subjective view. This helps in retrieving documents according to their linguistic terms definitions not to our definitions.
- Retrieve a set of relevant documents semantically using the proposed fuzzy ontology tool MVFRO.
- Deal with the multi-field topics problem using a predefined multi-view fuzz ontology.
- The resulted set of documents is ranked 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 our proposed fuzzy ontology tool.

References