A Non Trivial Approach for Ontology based Query Expansion Using Scoring

Misbah Fatima and Maruf Pasha

Department of Information Technology, Bahauddin Zakariya University, Multan, Pakistan.

Address For Correspondence:
Dr. Maruf Pasha, Department of Information Technology, Faculty of Science Bahauddin Zakariya University, Multan, Pakistan.
Phone No: 92 346 7111731; E-mail: maruf.pasha@bzu.edu.pk

ARTICLE INFO
Article history:
Received 18 July 2016
Accepted 21 August 2016
Published 3 September 2016

Keywords:
IR (Information Retrieval), WSD (Word Sense Disambiguation), W3C (World Wide Web Consortium), SW (Semantic Web), QE (Query Expansion), Ontologies

ABSTRACT
Background: The process of expansion of query is to reformulate the queries for improving performance of various operations regarding information retrieval.
Objective: Query Expansion (QE) technique is a most suitable technique to evaluate user’s input (i.e. words written in search query, sometimes-different keywords are typed) and to optimize the search results by associating some additional parameters.
The basic objective of this research is to propose a Query Expansion technique that utilizes ontologies for expansion. We have identified the various limitations in existing approaches and have tried to overcome the shortcoming of these techniques in our proposed architecture.
Method: The paper presents a layered approach for query expansion by associating semantic annotations and assigning weights to those concepts for query retrieval with precision.
Results: A sample case study was implemented to evaluate the working of the proposed architecture of query expansion. Results show that precision and recall was improved by adding semantic annotations and adding scores based on their relevance.
Conclusion: The proposed architecture provides improvement in query expansion results by evaluating the semantic relevance among words based on their weights using ontologies. This advancement in the domain of software and ontologies will play an indispensable role in QEs process and will also make a very clear, unambiguous impact on search results.

INTRODUCTION

Current Web holds a huge amount of data, which is available in the form of web documents. Retrieving data from such a large repository is a cumbersome process and requires a lot of effort to find out the precise information. Query Expansion is a technique mostly used by search engines to find out the user's required results by finding the relevance of keywords present in entered. However, the expansion of queries can be elaborated by adding number of additional words or phrases that frequently co-occur with the original query words. For example a User Query: “Toyota”; when expanded results in another query that is: “Toyota automobiles autos” etc.

Semantics are the senses, meanings, and explanation of words, symbols and complete structure of the sentence. The central goal in this study is to achieve the high expansion performance using Ontologies, and the Ontologies are highly related to the area of SW. Other main objective of the study is to achieve and to touch the more related data with the user query. And to improve the retrieval of information and the query response time for the user query very short. The proposed system is related to the field of the semantic web for the QE by using indexing and the Ontologies. The concept of Ontologies by the different scholars is; in 1993, Gruber at first distinct the details of an Ontology as an “An explicit specification of a conceptualization.”i In 1997, Borst defined Ontology as: “A formal and explicit specification of a shared conceptualization”ii
Furthermore these explanations of definitions that are most wanted that the conceptualization let to express a shared outlook bordered by lot of various parties, a common agreement instead of person to person analysis. Further, such a conceptualization (official) should come into view in devices readable form. In 1998, Studerte al. combines these two definitions stating that: “An Ontology is a formal and explicit specification of a shared conceptualization.”

Related Work:

Technique of query expansion is used to improve the performance of retrieval by adding the most related additional words, synonyms, concept matched and description of the words with the user original query. A QE concept has been proven a remarkable technique to get the useful and precise results. QE can be generated manually or automatically by the computer system with help database. It should be done by adding some words statically not randomly generated by the system or dynamically matched with the concept of the original query. Some techniques used the relations and some used the approaches of the QEs. Relations of the QEs are synonyms, antonyms, polysemy, monosemous, and the approaches are Local analysis of the query and Global

By examining the currently used QE techniques, it was observed that most of the techniques make use of WSD, and queries of the user’s can be executed by WordNet. WordNet is the dictionary of the words, synonyms, antonyms, all relations of the QE handled in this dictionary, it only process the simple queries not effective for the complex queries (Qiu, 1993), statistical model for the relevance and for the expansion (Navigi te al., 2003), ontology (Wan te al., 2012) (Snow te al., 2006) (Carpinetto te al., 2012) (Hirst, 2009), use of indices (Sharifullah te al., 2013), Global and the Local analysis (Chandrasekaran te al., 1999), ontology matching (Xu te al., 1996) (Andreu, 2005), concept to query closeness (Wan te al., 2012), use of probabilities for hyponyms acquisition (Zhang te al., 2008), automatic QE (Alejandra, 2012), QE effectiveness (Carstens, 2011), query log (Ben te al., 2009), knowledge-based retrieval (Lopes te al., 2014), information retrieval (Liu te al., 2005), Semantic based Expansion of Query and fuzzy ontology for query refinement (Alfred te al., 2014) (Shabanzadeh te al., 2010) (Dwi, 2001), concept based QE for multimedia retrieval (Takagi te al., 2001), information extraction from the database and retrieval for expansion to the user, this approach gave good and the precise results in the precision but still having some ambiguities (Stojanovic te al., 2004). In the past the QE was done manually or by the computer generated thesaurus, significance response, mechanically or statistical QE or by interactive QE. By scrutinize to be had techniques of the QE, these techniques make exploit of the NLP based elaboration of query, the statistical model (Sharifullah te al., 2013), Ontology (Chandrasekaran te al., 1999), knowledge of Semantic (Xu te al., 1996), knowledge based [Jing te al., 2012], SW in the case of web queries (Zhang te al., 2008), compatibility of the user queries (Alejandra, 2012) and Conceptual based QE (Snow te al., 2006), user logs (Carpinetto te al., 2012) (Carstens, 2011) (Guarino te al., 2009), co-occurrence (Ben te al., 2009), viewpoint orientated exploitation (Hirst, 2009). The user query compatibility based on fuzzy set of laws (Lopes te al., 2014) and different statistical models. The compatibility among the elaborated phrases and the existing files in the thesaurus can be intended by the method of phrases of co-occurrence (Ben te al., 2009), Google resemblance gap (Andreu, 2005). Semantic measures of the publications in scientific domain have been evaluated using citation networks. (Balaji te al., 2016)

Techniques of the QE are sufficient to deal with the different troubles like mismatching of the vocabulary or gaps of the vocabulary (Liu te al., 2005). The basic concept of the gap of the vocabulary is, dissimilarity among commented on idea or the conception and the query of the user, because WSD is capability method on the way to search out the sense of sentence expression in its different framework (Alfred te al., 2014). The researchers are paying attention to elaborate the queries of the user by using different resources of lexical like that the lexical relations of the semantic (Shabanzadeh te al., 2010), co-occurrence of the word’s frequencies (Dwi, 2001), WordNet (Qiu, 1993). WordNet chain of command and synsets have also been used to perk up the Wikipedia categorization (Takagi te al., 2001). WordNet is also useful for method of Conceptual QE (Stojanovic te al., 2004). The WSD of the client queries be able to act upon by using WordNet. The dictionaries are like some resources for example WordNet has increased a considerable investigator’s concentration and produces authentically creative outcomes for the very trouble-free queries. Although the WordNet based elaboration of query it generates fine outcomes for the squat and uncomplicated queries but it is not useful for difficult or semantic supported queries basically it is just obsolete for that semantic type of queries (Apostol te al., 2007). The WordNet elaboration relies only on the senses of the lexical rather than the abstract sense of the query. Based on the concept mash query is still an interesting problem. Semantic questions are based on the combination of different materials, actions and scenes also the semantics. Sense information of everyday life, the corporal, social, and psychosomatic phase of life, including. To get information from numerous research studies shows the significance of understanding the reasoning, data mining, and removal of noise means filtration of data etc. and finding polarity through semantic similarity (Zhanjun te al., 2006) (Sangeeta te al., 2015).
This amendment restricts the user to deal with the question of the need for semantic expansion. The proposed technique lexically also user query semantically make expansion. Lexically the user query will be extended and elaborated by the using of WordNet. But the expansion or the elaboration of the query that is based on the semantics it will be completed by using Concept Net.

**Proposed Architecture:**
Flow of the proposed system is when the client enter query for finding some specific concept. The given input from the user may be in incorrect language (English). The proposed system will be working by follow various process, in the start user will send his query/problem to the problem query here the different process will done like first of all syntax analysis will analyze the query and modified the query in proper format after that semantic analysis, analyze the problem query for semantic checking, after the completion of these two process further the query problem will move toward parse query here the query problem will divided into chunks, further steps which will be followed by problem query explained in below proposed architecture of the system. Fig. 1 shows the complete steps of the proposed architecture.

**Architectural Components:**

**A. Problem Query:**
The processor will execute the queries that were given as input by user and the processor will eliminate the noisy words from the inputted sentences and also allot the index number to the extracted meaningful words from the query it is all via a processor. The processor will send the complete information to the step of the proposed system named by query interface. And the processor assign as keywords to the extracted words from the user query. These keywords based on the weight, closeness and the distance from the main concept of ontology, which is matched with the user, input query. For the keywords, in list all meaningful and extracted words after the removal of the noise as a keyword. These in listed and extracted keywords match the matching conditions like sameAs, and exact matching of the words with the ontology. And the query processor eliminate all the preposition or connector words of the sentence like and, the, of, is, are, to etc. as noisy words and will never consider to these words as a keyword. And the query processor will consider meaningful word and those words that have their sense and also have synonyms and additional information in the list.

The client will input the query to the query processor named with query interface, and it will divide the complete sentence or the query into the parts or chunks. All will do for the keyword extraction. The stem words are already save in the ontology and the extracted list of keywords will match with the stem words of the ontology and the basic rules of the proposed system applies on the stemmers and keywords which will retrieve through the retrieval of the classes or subclasses of the system. All the words or the terms will retrieve matched words, equivalent stem words and will match the words by using the indexing methods or the techniques.

**B. Query Analysis:**
All the extracted information of the problem query will move to the Query analysis. After the removal of the noise the keyword list will be in the proper form. And this keyword list is useful for the query analysis and will be useful for finding the root word in stemmer and also the stem words are most appropriate for the extraction of the synonyms from the ontology server. For the extraction and the expansion of the keywords, or for extracted root words and also for the token generation against the keywords the Query analysis will connect directly with the server ontology.

C. **Query Hypothesis:**
A hypothesis is an assumption or intended to details complete on the source of partial evidence as a preliminary end for more exploration. In query hypothesis the query is the idea of query will map with the user query idea. All the retrieved information from the server and the query analysis will map with the user query idea.

D. **Server ontology:**
The server ontology contains the extra information and synonyms of the keywords. When query will operate, the user will obtain number of information through applying special sort of queries. User can’t moderator that is the most excellently relates and closed choice with his query. Then user will obtain more meaningful words, which fulfill the exact user query.

E. **Elaborate Query:**
In elaborate query process all the retrieved information that is in the expanded form will sum up in a single one query. By using, the mathematical function of the union ($\bigcup$) all retrieved information of all the previous component of the proposed system will sum up by this union function. Information of elaborate query step will sends back to the problem query step and the problem query set it in the proper format of the query or the sentence form for the user understandability and the problem query will be send back to the user with the additional word or expansion of the query and also added the additional information or the description of the different words.

**Experiments:**
As an example, a small case study was formulated and implemented to evaluate the results. A set of documents related to assign each document to solitary individual of collected two groups, these two groups are "appropriately related" documents and the 2nd is "not related" to the user demanded query. In this occurrence, all the appropriate "associated" documents are just only individuals that fit into the "related" group. Each group is then assigned a fixed weight for evaluation at a later stage.

**Table I:** Basic Rules Of Function And Scoring

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Function</th>
<th>Notation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>exactMatch()</td>
<td>$\alpha$</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>sameSet()</td>
<td>$\beta$</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>subClassOf()</td>
<td>$\gamma$</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>EquivalentOf()</td>
<td>$\omega$</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>disjointSet()</td>
<td>$\phi$</td>
<td>0.0</td>
</tr>
</tbody>
</table>

A detailed work rate, the particular and the explicit information for a solitary group of words will then summed up. Some concepts are here: Factual positive, bogus negative and the fake negative. In factual positives the number of objects that were extracted and the exactly matched with the user query fall in the factual positive concepts and deals it in the elaboration query. And bogus positives, extracted information are wrongly tagged in class. Inducing in this fact, it is different as the number of documents of the factual positive separated or divided by the all number of essentials that are actually place on the exact matched place that is the
sum of accurate positively and fake negative. And fake negatives, the extracted word placed on the right placed but unfortunately all these are not exactly matched or meaningful.

**Basic Rules For Matching And For Expansion:**

The architecture of the proposed system basically extracts the information from the lowest level of the ontology for the elaboration of the input query by using proposed algorithm all information extracted from server ontology. In simple words the basic rules for the QE are the practical example of the extraction of word from the ontology. The basic function of the Ontologies are exactMatch, sameAs, subClassesOf, EquivalentOf and the disjointOf, are built in methods used in the OWL. And we assigned scores of these functions or the levels of the ontology for the query expansion.

![Fig. 2: Case study for Teacher-Student ontology](image)

All the subClassOf, exactMatching, sameAs, EquivalentOf and disjointOf here the subClassOf method or all these functions are used for the elaboration of the user query in different ways because the extracted information is in the form of additional word, synonyms and to enhance the query. Function sameAs play a key role when user put his/her input query in the abbreviated forms of terms then the system understand the query and extract all those information from the ontology which is user required consequences. Subclassof and sameAs gives more extensive consequences in the context of the recall. But the exactMatching task provides out the précised and exact consequences and the full QE. disjointOf function gives out the result not in the expanded form. It might be possible that the word doesn’t exist in the vocabulary of the ontology or may not be a related words exist here. It is also may be no relationship defined or it may the user enter the combination of alphabets not enter the meaningful word. Fig 2. Shows the complete case study of implemented Teacher student ontology to highlight the similarity and equivalent concepts. Fig.3 shows the equivalent concepts that can be treated as sameAs. For example isEducating concept can also be referred as is Teaching and therefore should retrieve all possible results. After executing various sets of queries precision and recall factors were calculated as shown in Table 2 and graphically projected in Fig. 4.

![Fig. 3: Equivalent relations](image)

**Table II:** Precision & Recall Factor For Full Query Expansion
Conclusion:

The proposed system is categorized into a component-based approach where each component is highly cohesive and designated to perform only individual tasks. The user receives a number of expanded terms amongst which he can choose the most appropriate queries to choose from. The system will regain the complete information regarding the ontology classes and also regains the equivalent class. The expansion or the elaborations are fundamentally based on the equivalence classes. Results show that precision has tremendously improved by assigning scores to concepts. More precise results were found by using a layered approach. Synonyms, hypothesis and additional information in the query were used to find all the relevant word, weighted and same sense terms that also improved recall tremendously. In future we want to apply this semantic based search in web services to improve the service discovery process.

REFERENCES

Chandrasekaran, B., John R. Josephson, 1999. What Are Ontologies and Why Do We Need Them?”, IEEE.
Snow, R., D. Jurafsky, and A.Y. Ng, 2006. Semantic taxonomy induction from heterogeneous evidence, In ACL.