ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

EFFICIENT QUERY KEYWORD INTERPRETATION FOR SEMANTIC INFORMATION RETRIEVAL

Sonia Setia*, Jyoti Verma, Neelam Duhan

Department of Computer Engineering, J. C. Bose University of Science and Technology, YMCA, Faridabad, INDIA

ABSTRACT

Due to vast information available over the World Wide Web and its unstructured nature it is becoming more difficult to get relevant information. Currently, information retrieval techniques are keyword based. They do not finally capture the semantic meaning of a query keyword. To overcome this problem, conceptual knowledge to information retrieval has been introduced by means of taxonomy. In this paper, a semantic information retrieval mechanism has been presented which translates query keywords to categories belonging to a taxonomy. A hybrid similarity method has been proposed which finds the closest category corresponding to a keyword using a thesaurus like WordNet. Evaluation of the proposed approach has been done for semantic based declarative querying process which shows better results in terms of precision, recall and F-measure.

KEY WORDS
Information Retrieval, Semantic-based IR, Taxonomy, categorization, keyword-based IR

Received: 15 Jan 2020
Accepted: 21 Apr 2020
Published: 3 May 2020

***Corresponding Author**
Email:
setiasonia53@gmail.com
Tel.: +91-8383007704

INTRODUCTION

In past few years, keyword-based information retrieval systems have been used to retrieve information over World Wide Web. Currently, most search engines are purely based on keyword based information retrieval. They accept query in the form of keywords and output those documents which contain the given keywords. But these search engines do not consider the semantic meaning of those provided keywords. Therefore, they provide number of false links of documents and users are not able to find relevant information. The main aim of an information retrieval process is to retrieve the relevant information corresponding to the given query. In particular, this requires understanding users' needs precisely enough to allow for retrieving a precise answer using some semantic technologies. Taxonomies appear to be useful method to allow for more semantics based search. Therefore, there is a need of translating keyword-based information retrieval to category-based information retrieval. In this paper, an approach has been presented to interpret query keywords using knowledgebase available through taxonomies. Based on presumptions about how individuals portray their data needs, proposed approach translates a keyword based query into category-based query. The evaluation of the proposed methodology has been done on queries given by few users at our institute. It uses the knowledge base of the semantic portal available at http://www.dmoz.org.in/ and displays better results in terms of precision, recall and F-measure.

A, significant work has been performed in literature to find the similarity between two elements. While these approaches claim for remarkable results, but the approach is not clear enough that how this has achieved. In fact, it is observed that users are more comfortable in keyword based search. But, it also seems important to design an approach for interpretation of keywords such that more meaningful and relevant information can be retrieved.

Chahal et al. [1] proposed a technique to compute similarity for semantic web documents that is based upon conceptual instances found between the keywords and their relationships. Authors explored all relevant relations that may exist between the keywords which explores the user's interest and based upon that determine the similarity between documents.

Formica [2] proposed a similarity measure for Fuzzy Formal Concept Analysis (FFCA), which is a general form of Formal Concept Analysis (FCA) which is used for modelling of uncertainty information. Although FFCA became very popular for semantic web development. But the problem with the given work is that manual development of ontologies is a time consuming process. Further, for constructing the fuzzy ontologies Zhang et al. [3] proposed an automated approach by using Fuzzy Object Oriented Database (FOOD) model. This way it supported the automated process for retrieval of information.

De Maio et al. [4] proposed new retrieval approach which is based on ontologies. By supporting data organization and visualization, it provides a friendly navigation model. The major challenges faced by researchers are to find the efficient techniques of sharing and searching the information with the rapid growth of web. By using the concept of Fuzzy, Kohli and Gupta [5] solved the challenges of information retrieval system. Aloui et al. [6] proposed a semiautomatic method to design and extract ontology which is based on clustering, fuzzy logic, and formal concept analysis (FCA). Authors represented the ontology as a set of fuzzy rules. Prot´eg´e 4.3 has been used to evaluate the proposed approach. Results shows that by using ontology mapping, more relevant information can be retrieved. Kandpal et al. [7] proposed a new methodology for ontology alignment. Ontology alignment is done by retrieving the similar concepts of two different ontologies. If directly concepts are not matched of two different ontologies, then similarity can be calculated of expanded terms. Major challenge is to provide accurate information of user's uncertain query words.

IIOAB JOURNAL
ISSN: 0976-3104

Rani et al. [8] proposed a hybrid retrieval system which integrates ontology and fuzzy logic concept to find information. Fuzzy type 1 has been used for documents and fuzzy type 2 has been used for words to prioritize the retrieved list.

A critical look at the literature motivates us to propose an approach to retrieve more relevant information by considering the semantics of user's query. Here, in this work, we have proposed a method which interprets the query keywords semantically in the form of taxonomy. To achieve this task, keyword to category (terms belonging to taxonomy) mapping has been done by using proposed hybrid similarity matching method. Currently, no real automatic solution has been found using knowledge-base for keyword to category mapping. So, the surveying of literature work motivates the semantic mapping of keywords by using the concept of taxonomy to retrieve more relevant information.

MATERIALS AND METHODS

This paper proposed a taxonomy-based approach for query interpretation which is on the ambition of producing more precise query from a given keyword so that more relevant information can be retrieved. Domain taxonomy has been used to retrieve more precise query in the form of categories belonging to taxonomy. Therefore, a keyword to category mapping approach has been proposed which is depicted in [Fig. 1].

1. At the first stage, users put queries for retrieving information, these queries are parsed into keywords or phrases, typically n-grams (n-gram is a n word sequence).
2. To retrieve more relevant information, these n-grams are mapped to the categories T = {c1, ..., c¬k.} of a domain taxonomy. This mapping is performed using a similarity matching method which is based on domain specific taxonomy. It uses thesaurus to find closest category corresponding to a keyword. Moreover, we used WordNet as thesaurus.
3. To better characterize the objects, weights are assigned to the keywords according to the frequency of queries corresponding to an object. Therefore, the taxonomy categories' weights are also updated. And finally, the resulted category based query is passed to Search engine for more relevant information retrieval in terms of this precise query.

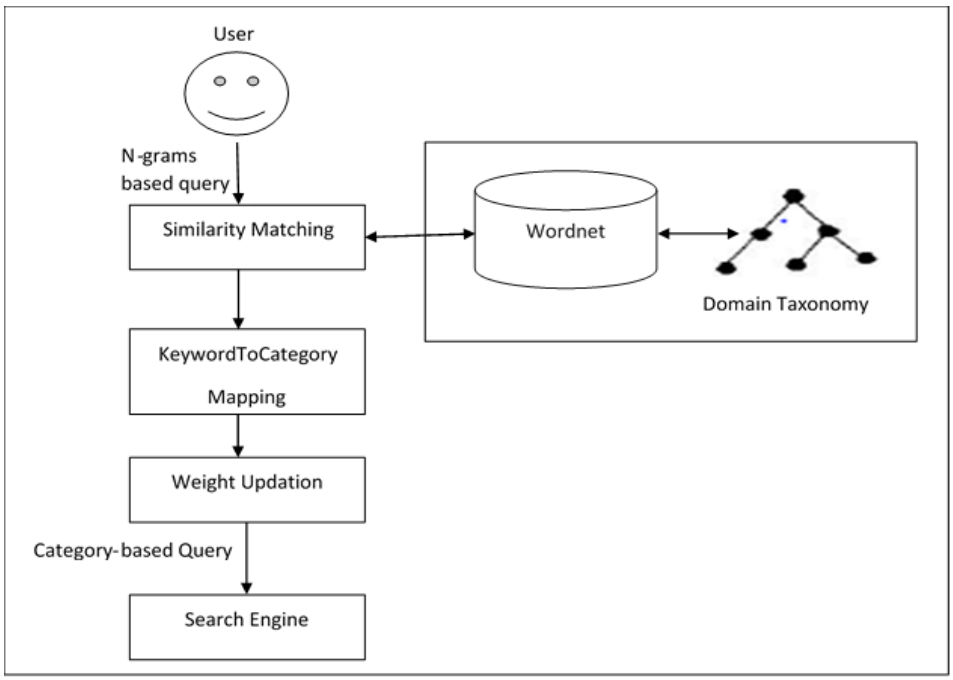

Fig. 1: Translation of keyword-based query to category-based query

Wordnet

WordNet is an online lexical reference system where, English nouns, verbs, adjectives, and adverbs are organized into Synsets. Each one is representing an underlying lexical concept. "Synsets" is a set of synonyms which represent a concept or a knowledge of a set of terms. Synsets make diverse semantic relations for instance synonymy (similar) and antonymy (opposite), hypernymy (super concept)/hyponymy (sub concept) (also known as a hierarchy/taxonomy), meronymy (part-of), and holonymy (has-a). For each keyword in WordNet, we can have a set of senses. For example, the word "wind" has eight verb senses and eight noun senses. The first sense of "wind" as a noun gives the following path:

ISSN: 0976-3104

wind –>weather; weather condition,
atmospheric condition→
atmospheric phenomenon-->
physical phenomenon--> natural phenomenon,
nature--> phenomenon

Similarity matching method

It is the major component for our semantic retrieval mechanism. Query interpretation is done by using similarity matching method. In order to find the closest category in the taxonomy T for a keyword k, we calculate the similarity through the mechanisms provided by the thesaurus.

If the keyword belongs to the taxonomy, then it is included as it is. Otherwise, most similar category is found corresponding to the keyword by using proposed Hybrid similarity method which is the integration of Type-based similarity and Path-based similarity.

Type-based similarity: If a keyword k has been defined as synonym of a category c it means keyword is directly related to this category i.e. keyword is a type of this category. Then this category is assigned to the corresponding keyword with similarity value 1 and no need to compute similarity with other categories.

Path-based similarity: If no direct synonym is found in that case, path based similarity will be computed for keyword to category mapping. Wu & Palmer similarity measure has been used to compute similarity between senses of k, Sn(k) and the categories c in T, that measures similarity between two terms. We select the pair (k,c) which is having maximum similarity and map keyword k to the taxonomy category c.

Using Wu & Palmer similarity we can compute the path-based similarity between two nodes a, b of the given taxonomy by using following formula:

$$S(a,b) = \frac{2 * \text{depth}(LCS(a,b))}{\text{depth}(a) + \text{depth}(b)}$$

where LCS is Least Common Sequence of a and b.

Breadth First Search traversal algorithm has been used to traverse the taxonomy while comparing the keywords with categories to reduce the search space. First it compares the keywords with the categories at top level of taxonomy. The category having highest similarity is explored further to find relevant sub-category. This process is repeated until most relevant category is found. Finally, keyword and category pair i.e. (k,c) pair that gives maximum similarity s has been selected. After this complete process, each keyword is mapped to a category with a similarity s respectively. Once a query has been augmented with appropriate categories it can be handed over to a search engine that is designed to pinpoint information. The challenge of the algorithm is to be able to select the right category corresponding to keywords in order to improve the information retrieval. The algorithm follows these steps:

```
Algorithm keyword Category Mapping(k, t)
1. For all sns ϵ Sn(k) do
2. For all c ϵ T do
3. sim ← max(WPsim(sns, c));
4. done
5. snscsim= max({sim});
6. cmax = c ϵ O, for which (sim == snstsim);
7. done
8. kcsim = max({snscsim});
9. category = c ϵ {cmax}, for which (sim == kcsim);
10. return(category, sim);
11. done
```

Algorithm analysis are broadly classified into three types such as.

- Best Case: If keyword is available in taxonomy then it is considered as it is for information retrieval.
- Average Case: Otherwise best possible match is found for a category which has the maximum similarity with keyword.
- Worst Case: If there is no category found for a keyword then ignore that keyword and we assume that keyword doesn't belong to our specific domain.

v) Weightage of the Resulted query. The normalized weight has been assigned to every mapped category derived from the similarity matching algorithm which can be calculated by given formula
Weight of category= w*s

Where, w represents weight of the keyword
s represents similarity value between keyword and mapped category

IIOAB JOURNAL
ISSN: 0976-3104

RESULTS

The proposed approach for the translation of query keywords with respect to a domain specific taxonomy is incorporated in our prediction system framework called Semantic Prefetching System [9] which has been intended to help a blend of search and investigation in information bases. We will presently depict a potential interaction of a user with the proposed framework. For the evaluation of the proposed approach, we have asked our colleagues at our institute to provide queries. It uses the knowledge base of the semantic portal of Dmoz [10]. Few of them were expelled which were out of scope of our domain specific knowledge base. For the evaluation, users manually allotted conjunctive queries corresponding to the natural user queries. A query produced by our approach is considered as accurate if it recovered indistinguishable answers from the hand crafted query. Few examples of the queries given by our users are shown in [Table 1].

Table 1: Translation of query to conjunctive query

User Query	Corresponding Conjunctive Query
Guitar	Stringed instrument
Techno	Dance
Karaoke	Music equipment
Veena	Stringed instrument
Flute, Sitar	Wind Instrument, Stringed instrument
Piano	Keyboard instrument
Vocoders	Electronic instrument

We evaluated the proposed approach in terms of precision, recall and F-Measure. Precision P is calculated by the number of accurately interpreted query keywords divided by the total query keywords interpreted by system. Recall R is calculated by the number of accurately interpreted query keywords divided by all the query keywords. F- measure is harmonic mean between precision and recall. In case, the query is interpreted automatically by our system, our system obtains a precision 84%, a recall 72% and F-Measure 77% as depicted in [Fig.2].

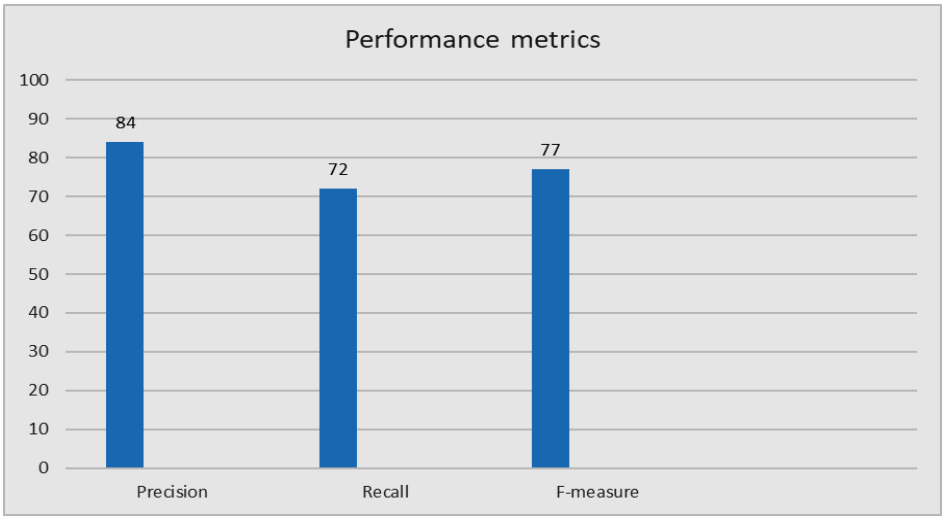

Fig. 2: Performance metrics for our proposed approach

DISCUSSION

The proposed approach has been evaluated with the knowledgebase of music domain [10]. Evaluation has been performed by our colleagues who do not have any knowledge of considered domain. They have been asked to provide general queries related to music domain. Some of the queries were obviously removed which were out of scope of music domain. Noticeably, this is a problem as this will affect the recall measure. But, here we have achieved 72% recall which means 72 out of 100 user given queries has been mapped to domain taxonomy terms which is a great achievement as compare to paper presented in [11], where, authors claimed 50% recall for their proposed approach. Precision shows that the generated conjunctive queries by our approach is correct in most of the cases which is approximately 84% which is also a large percent as compare to [11] where, authors claimed 69% precision. In short This paper proposed a novel approach for keyword to category mapping so that more precise query can be retrieved for better results. A novel hybrid similarity matching method has also been proposed which has been evaluated against few users given queries. Results shows that our approach gives 84% precision, 72% recall and 77 % F-measure. Novelty of the work has been covered in the following points:

IIOAB JOURNAL
ISSN: 0976-3104

- A novel similarity matching method between two words has been proposed, which uses Thesaurus like WordNet to find similarity. Breadth First search traversal algorithm has been used to traverse the taxonomy while calculating similarity. It is a hybrid approach which integrates Type based and path based similarity.
- A novel keyword to category mapping technique has been proposed which exploits the proposed similarity matching method to find similarity between user given query keyword and category belonging to taxonomy and finally, finds the best matched category corresponding to users given query

CONCLUSION

This paper proposed an approach for interpretation of query keywords in more precise manner. It supports the Information Retrieval system to overcome the limitations of traditional Information retrieval system so that users can retrieve more relevant information respective to their queries. It also helps to improve Hit-Miss ratio. By using domain specific taxonomy, proposed system will support information retrieval system semantically. This system can handle the semantic issues for information retrieval.
Based on the proposed approach, retrieval system can be extended to support different domains. It also can be extended to other local languages.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
I would like to express my special thanks of gratitude to my supervisors Dr. Jyoti and Dr. Neelam Duhan, it is truly an honor. Thank you for all the advice, ideas, moral support and patience in guiding me through this project.

FINANCIAL DISCLOSURE
There are no financial conflicts of interest to disclose.

REFERENCES

[1] Chahal P, Singh M, Kumar S. [2013] An ontology based approach for finding semantic similarity between web documents. International Journal of Current Engineering and Technology, 3(5): 1925–1931.
[2] Formica A. [2013] Similarity reasoning for the semantic web based on fuzzy concept lattices: an informal approach. Information Systems Frontiers, 15(3): 511–520.
[3] Zhang F, Ma Z M, Fan G, Wang X. [2010] Automatic fuzzy semantic web ontology learning from fuzzy object-oriented database model. Database and Expert Systems Applications, 6261: 16–30.
[4] DeMaio C, Fenza G, Loia V, Senatore S. [2012] Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis. Information Processing & Management, 48(3): 399– 418.
[5] Kohli S, Gupta A. [2014] A survey on web information retrieval inside fuzzy framework. In Proceedings of the Third International Conference on Soft Computing for Problem Solving, 259: 433– 445.
[6] Aloui A, Ayadi A, Grissa-Touzi A. [2014] A semi-automatic method to fuzzy-ontology design by using clustering and formal concept analysis. In Proceedings of the 6th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA '14): 19–25.
[7] Kandpal A, Goudar R, Chauhan R, Garg S, Joshi K. [2014] Effective ontology alignment: an approach for resolving the ontology heterogeneity problem for semantic information retrieval. Intelligent Computing, Networking, and Informatics, 243: 1077–1087.
[8] Rani M, Muyeba M, Vyas O. [2014] A hybrid approach using ontology similarity and fuzzy logic for semantic question answering. Advanced Computing, Networking and Informatics, 1: 601–609.
[9] Setia S, Jyoti, Duhan N. [2019] Semantic Prefetching Based Hybrid Prediction Model. International Journal of Scientific & Technology Research, 8(12): 3936-3941.
[10] http://www.dmoz.org.in
[11] Tran T, Cimiano P, Rudolph S, Studer R. [2007] Ontology-Based Interpretation of Keywords for Semantic Search. ISWC/ASWC, LNCS 4825: 523–536.