Asian Journal of Information Technology 17 (3): 197-201, 2018
ISSN: 1682-3915

Ontology Based Efficient Multi Keyword Query Interface for Search Engines

[1]S. Jayasundar, [2]V.N. Rajavarman and [2]V. Saishanmuga Raja
[1]M.G.R Educational and Research Institute, Chennai, India
[2]Shanmuganathan Engineering College, Pudukottai, India

Abstract: Keyword search is an efficient data retrieval method for the WWW, largely because the simple and efficient nature of keyword processing allows a large amount of information to be searched with fast response. However, keyword search methods do not formally capture the clear meaning of a keyword query and fail to address the semantic relationships between keywords. As a result, the accuracy (precision and recall rate) is often unsatisfactory and the ranking algorithms fail to properly reflect the semantic relevance of keywords. Our research particularly focuses on increasing the accuracy of search results for multi-word search. We propose a statistical ontology-based semantic ranking algorithm based on sentence units and a new type of query interface including wildcards.

Key words: Wildcards, ranking, semantic, algorithm, query, address

INTRODUCTION

Presently the keyword searching is efficient in the searching methodology due its high efficiency. But it does not provide a semantic understanding of the keywords because it is difficult to find the exact meaning of the keyword without considering the semantic relations of the word or without knowing the full context of the sentence. At the same time the search results are not convincing. When a user is searching some information in the search engine if the information that is being searched is not highly ranked then the user may search the information again and again with a new query rather than clicking through the next pages. This happens because the existing ranking algorithms do not map the semantic relevance between the query and the web contents. In this study, we introduce a new query interface which keeps one or more tags between keywords or at the beginning or at the end of a query. This will allow search engines to return exactly what the user is searching in an efficient way. For example, if a user searches about the price of a car then the user has to place a query of price (tag), car. This new query interface calculates the frequency of occurrence of the keyword in the position of tag as relevant to actually what the user is looking for.

The main objective of the research is to increase the accuracy of search results measured by means of recall rate and precision. For this, we propose a new query interface having a tag and ontology based semantic ranking. In first phase, we provide high ranking to the keywords present in same sentence rather than the keywords in different sentences. While existing statistical search algorithms such as N-gram (Rosenfeld, 2000) only consider sequences of adjacent keywords our approach is able to calculate sequences of non-adjacent keywords as well as adjacent keywords. In the second phase, we propose a query interface which considers the tag as an independent token of a search query to relate to what actually the user is searching. Unlike the existing information retrieval approaches such as proximity approaches, semantic and natural language assisted search approaches (Fernandez *et al*., 2011; Ruiz-Casado *et al*., 2007) statistical language modelling, query prediction and query answering our approach helps in improving information retrieval efficiently.

Literature review: The most important factors which current search engines, including Google, adopt to determine their ranking results for multi-keyword search are frequency and proximity (Jansen *et al*., 2000). One of the main problems with the current ranking algorithm of multi-word search arises from the fact that its methodology calculates the relevance of keywords only by their proximity without considering whether they exist in the same sentence or not. For this reason, this method fails to consider the possibility that multiple neighbouring keywords have no relevance to each other for example when one word is placed at the end of the first sentence and the other in the beginning of the second sentence. Another problem is that even when multiple keywords are semantically closely relevant, if other words such as modifiers are inserted between them the current ranking methodology calculates their relevancy as low. The other problem is that this methodology cannot successfully recognize semantic differences between

Corresponding Author: S. Jayasundar, Dr. M.G.R Educational and Research Institute, Chennai, India

multiple keywords whose orders are reversed for instance, dog eat and eat dog. To overcome these problems, the following technologies have been introduced.

While the exact details of how current search engines perform their indexing and rank query results are kept mostly secret for competitive reasons and to prevent manipulation by end users it is generally understood that crawlers are fed numerous seed URLs and tokenize the text in the web pages they find to be analyzed, following links they find and then tokenizing the text on those pages to be analyzed. Another problem is that existing search algorithms fail to capture the semantic relevance of sequences of keywords when they are not situated adjacently due to an insertion of other words not included in the query such asadverbs or adjectives.

Ontology based search model: The ontology based search model adopts a new form of query interface which is shown in Fig. 1. A new html page containing three input text boxes are designed to get input keywords from the user. The user must enter one keyword each in the two text boxes. The one text box is left empty which is used to enter the tag to predict the actually what the user wants to search.

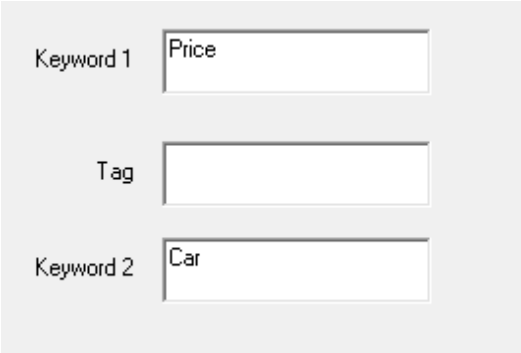

Fig. 1: Ontology based search model

The new query interface allows the user to enter the keyword they are looking for in the form of tag. Mostly the users do not know what they are exactly searching for but they may know only some keywords related to the search. Most of the existing search engines does not handle this properly because they provide search results based on the user inputs without knowing the semantic relevance of the search. Our approach tackles this problem by using the tag. This new query interface using the tag significantly reduces the volume of data to build the ontology. Moreover, our approach is able to return the most frequently used keywords in the location of the tag from the actual web data. This method can also research as a query prediction system.

In order to test our hypothesis referring to the ontology (Wyssusek, 2006; Chung *et al*., 2006) and adopting a new query interface produce more semantically relevant search rankings, we have developed a statistical ontology-based semantic search model. Using the model, we have built the ontology and created a new query interface which have generated and re-ranked a query-relevant subset of the corpus of English news text from the TREC (Voorhees, 2001) with a list of questions and answers for each query. The process of the new query interface is shown in Fig. 2.

As shown in the Fig. 2, the new query interface makes use of the web documents called corpus. The statistical ontology-based semantic search has been built

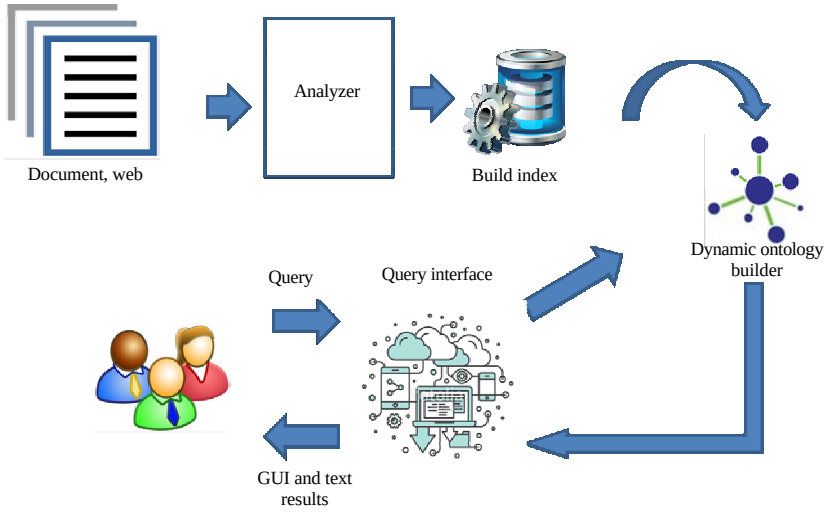

Fig. 2: Architecture of ontology based ranking

referring to the index structure of search engines. To perform the indexing of our data, we used and modified the open source Apache Lucene Indexer (Version 2.9.1) (Anonymous, 2017) and pointed it to look at all of our individual TREC documents. We began by using the built-in stop words analyzer while modifying the white space tokenizer and filter. Whereas existing search engines remove sentence delimiters while indexing, so that, they are not able to process data by sentence unit our approach allows the tokenizer to discard all symbols other than sentence delimiters such as periods as all items are one character in length. In addition to removing tags, whitespace and "stop words" such as "and" "the" and "to" we added "www", "http", "copyright" and other words that appear frequently in the footers of web pages and that are believed to be unnecessary when looking at the domain of our corpus. Users can modify a text file to add/remove stop words or specify ones in addition to the defaults in the command line. In this way, users can change the list of stop words for various types of corpus.

Experimental data: Even though our statistical ontology based search model is mainly designed for web data retrieval, the TREC which is off line corpus is used for quantized evaluation of search performance. It is because the TREC data supplies questions and correct answers and this helps us to show the effectiveness of our search model in a quantitative way. In detail, the TREC data which we have used, comprises approximately 2.5 GB of text (about 907K documents) covering the time period of October 2004-March 2006. In the TREC data, the contents will include people organizations, events and other entities. Each question series is made up of some factoid and some list questions. Factoid-type questions only have one single answer. Therefore, this type of question does not allow us to easily evaluate the recall rate of search results whereas list-type questions with multiple correct answers make the task easier. For the experiment, we calculated the precision and the recall rate by using twenty-two list-type questions.

MATERIALS AND METHODS

For evaluating our approach we make use of the following evaluation criteria.

Precision rate: In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

$$\text{Precision} = \frac{|\{\text{Relevant documents}\} \cap \{\text{Retreived documents}\}|}{|\{\text{Retreived documents}\}|}$$

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

$$\text{Recall} = \frac{|\{\text{Relevant documents}\} \cap \{\text{Retreived documents}\}|}{|\{\text{Relevant documents}\}|}$$

F-score: A measure that combines precision and recall is the harmonic mean of precision and recall the traditional F-measure or balanced F-score:

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

For the evaluation of our model, we compare our statistical ontology-based semantic search to Google's desktop search and an open source search engine, (Anonymous, 2014). The precision rates and recall rates of each approach are compared. We chose Google desktop search over Google's web search because there is no way to force Google web search to crawl and index our whole experiment data whereas we can with Google desktop search. Even though Google's desktop search does not use page rank algorithm (Page *et al*., 1999) for ranking which is one of Google's major Web search algorithms, Google desktop search can produce the same result for our experiment as the experiment result we would expect with Google web search because the TREC data does not have hyper-links required for page-rank algorithm. Since, the page rank algorithm of the Google web search can be used in addition to our statistical ontology-based algorithms, the fact that our experiment is not concerned with the usages of page link algorithms does not undermine the possibility that our approach can improve current search engine's technologies.

The TREC news data consists of about 1,000 news files and each news file has about 1.000 articles. The correctness of our search results for the questions which the TREC data supplies is evaluated by considering whether our search results generate links to study which contain correct answers. Therefore, we created an HTML file for each study and made three search engines involved with our experiment index about 10000 HTML files in total.

Our experiment mainly dealt with a two-word query because a two-word query is the most common form of a query employed by users (Jansen *et al*., 2000; Mittal *et al*., 2004). We have selected two semantically most important words from each question on the list provided by the TREC data. We have applied the two-word queries to Nutch to our statistical ontology

based search model and to Google desktop search. We have evaluated the precision and the recall rate of the top ten search results produced by the search engines because most users look only at the first page of results and usually the first page produces 10 results. And then, we have produced final evaluation scores calculating the precision rates and recall rates together following the evaluation method offered by the TREC.

RESULTS AND DISCUSSION

In order to test both recall rates and precision rates of each search approach, we have used questions offered by the TREC data which have multiple correct answers. A total of twenty-two queries are used after excluding queries which have 0 results for all three search approaches. Figure 3 shows a result comparing the precision rates of each search approach. The X-axis shows the TREC's query ID and the Y axis is the precision rate (Table 1).

Our approach considering whether or not keywords are placed in the same sentence, adds one more constraint to the search conditions of the previous search algorithms, thereby filtering more irrelevant search results than original Nutch can. For this reason, our approach produces more correct search results, so that, its precision rate is expected to increase more than original Nutch. Meanwhile, Google desktop search showed a very high precision rate because Google desktop search is sensitive to verb tenses or conjugation during the search process. Hence, we came to know that Google desktop search is more focused on producing precise search results than retrieving a wide range of target corpus (Table 2 and Fig. 4).

As expected, calculating whether keywords are placed in the sentence or not helped our ontology based model to return a lesser number of search results than just calculating frequency and distance. When the number of search results decreases, its recall rate is expected to decrease. In our experiment our statistical ontology-based search's recall rate also has decreased by about 7% over original Nutch's recall rate. Meanwhile, Google desktop search shows a lower recall rate compared with our model and Nutch. This result can be easily expected because Google desktop search showed a higher precision rate in the previous experiment. This result demonstrates that Google desktop search considers more constraints during the search process in order to acquire a higher precision search result than our model and Nutch's.

In order to properly evaluate search engines, both precision and recall rate are generally considered all together. Our experiment also has evaluated search results this way. In order to evaluate three search approaches in

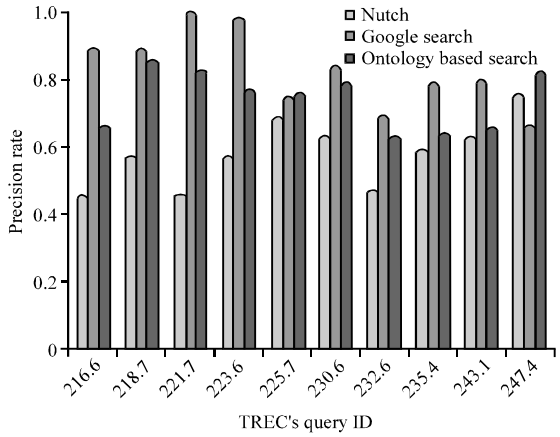

Fig. 3: Comparison of precision rate

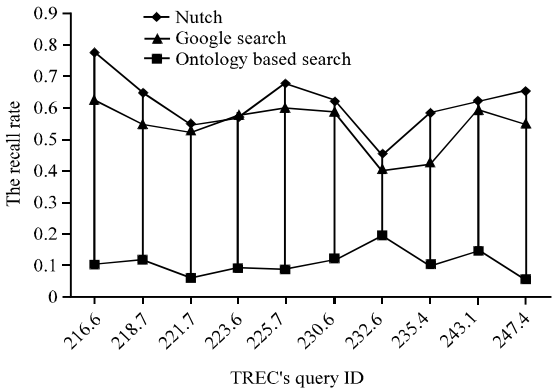

Fig. 4: Comparison of recall rate

Table 1: The precision rates for each query

TREC's query ID	Nutch	Google search	Ontology based search
216.6	0.450	0.888	0.6560
218.7	0.565	0.889	0.8540
221.7	0.454	0.996	0.8240
223.6	0.568	0.980	0.7654
225.7	0.684	0.745	0.7580
230.6	0.621	0.840	0.7840
232.6	0.459	0.691	0.6220
235.4	0.585	0.789	0.6321
243.1	0.623	0.796	0.6480
247.4	0.752	0.654	0.8450

Table 2: The recall rates for each query

Query ID	Nutch	Google search	Ontology based search
216.6	0.780	0.110	0.621
218.7	0.654	0.123	0.540
221.7	0.548	0.058	0.525
223.6	0.568	0.098	0.585
225.7	0.684	0.080	0.597
230.6	0.621	0.120	0.589
232.6	0.459	0.200	0.398
235.4	0.585	0.098	0.421
243.1	0.623	0.140	0.589
247.4	0.652	0.054	0.545

a quantized way and to consider both precision and recall rate equally, we have adopted a standard which the TREC suggests as shown below to aggregate recall rates and precision of each search approach.

CONCLUSION

In this study, we have presented a new multi keyword query interface using statistical ontology for improving the search rankings in search engines. By introducing higher-ranking scores to keywords in the same sentence for multi-word search our approach is able to produce more semantically relevant search rankings in the top ranked documents than Nutch and Google desktop search. From the experimental results it is also evident that by placing wild cards between keywords or at the beginning or at the end of a multi-word query, the new query interface helps to understand user's information demand more clearly. As a result with our approach, more precise and efficient information retrieval is possible. Furthermore, our ontology-based query interface which adopts a statistical language model for multi-keyword search, helps to generate semantically more relevant information retrieval results.

REFERENCES

Anonymous, 2014. Nutch, open source search software. Apache Software Foundation, Forest Hill, Maryland, USA. http://lucene.apache.org/nutch/.

Anonymous, 2017. Apache lucene and solr set the standard for search and indexing performance. Apache Software Foundation, Forest Hill, Maryland, USA. http://lucene.apache.org/.

Chung, S., J. Jun and D. McLeod, 2006. A Web-Based Novel Term Similarity Framework for Ontology Learning. In: On the Move to Meaningful Internet Systems, Meersman, R. and Z. Tari (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-48287-1, pp: 1092-1109.

Fernandez, M., Z. Zhang, V. Lopez, V. Uren and E. Motta, 2011. Ontology augmentation: Combining semantic web and text resources. Proceedings of the 6th International Conference on Knowledge Capture, June 26-29, 2011, ACM, New York, USA., ISBN:978-1-4503-0396-5, pp: 9-16.

Jansen, B.J., A. Spink and T. Saracevic, 2000. Real life, real users and real needs: A study and analysis of user queries on the web. Inf. Process. Manage., 36: 207-227.

Mittal, V., S. Baluja and M. Sahami, 2004. Google tutorial on web information retrieval. Sinai Hospital: Rubin Institute for Advanced Orthopedics, Baltimore, Maryland.

Page, L., S. Brin, R. Motwani and T. Winograd, 1999. The pagerank citation ranking: Bringing order to the web. Technical Report Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/.

Rosenfeld, R., 2000. Two decades of statistical language modeling: Where do we go from here?. Proc. IEEE., 88: 1270-1278.

Ruiz-Casado, M., E. Alfonseca and P. Castells, 2007. Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from Wikipedia. Data Knowl. Eng., 61: 484-499.

Voorhees, E.M., 2001. The TREC question answering track. Nat. Lang. Eng., 7: 361-378.

Wyssusek, B., 2006. On ontological foundations of conceptual modelling. Scand. J. Inf. Syst., 18: 1-19.