A Context Aware Approach to
Semantic Query Expansion

Najmeh Ahmadian
Department of Computer Science & Engineering.
Islamic Azad University- Branch
Dehaghan
Dehaghan, Iran
Najmeh.ahmadian@gmail.com

M. A Nematbakhsh
Department of Computer Science & Engineering.
Isfahan University
Isfahan, Iran
mnematbakhsh@yahoo.com

Hamed Vahdat-Nejad
Department of Computer Science & Engineering.
Islamic Azad University- Branch
Dehaghan
Dehaghan, Iran
hamed.vh@gmail.com

Abstract- Keyword based search scheme imposes the problem of representing a lot of web pages in the search engines. Query expansion with relevant words increases the performance of search engines, but finding and using the relevant words is an open problem. In this research we describe a new model for query expansion which employs user context and semantic concepts to discover new words for obtaining accurate results. Experimental results show an enhancement on information retrieval performance comparing to the traditional approaches.

Keywords: Information retrieval, Query expansion, Context, Semantic web, Ontology.

I. INTRODUCTION

Traditional search engines perform a keyword-based search scheme which leads to the problem of retrieving numerous irrelevant items. This scheme is very sensitive to the selected keywords. For example search with a specific keyword may eventuate to unsatisfactory, but with its synonym to appropriate results.

One solution to increase the relevancy of retrieved results is to expand the user query with relevant words. But finding appropriate relevant words is a challenging problem. In addition, sometimes the query is vague, in the meaning that its domain or context is not recognized and as a result, selection of augmenting keywords is really difficult. Instances of these queries are words which have various meanings in different domains.

According to research accomplished by Imaee and colleagues [1], query expansion has a significant role in the

Information retrieval. They showed that it can eventuate to the increase of precision in information retrieval.

Since the emergence of Semantic Web in late 1990, semantic search has been one of the most distinguished areas. Strategies for combining information of ontology [2] and electronic dictionary with search patterns are studied in several papers [3,4,5]. They show that exploiting only one semantic relation, such as Hypernym or Synonym is not effective, so it is better that a combination of semantic relations to be used. In the studies that a combination of these relations has been used [6, 7], promising results have been reported.

The contribution of this research is in two directions. Firstly, we intend to replace keywords with concepts and perform search in the semantic space. For this, we make use of ontology to define concepts and relations between them. Moreover, we utilize the context of the user to find their intention of the query. Context is generally defined as “any information that can be used to characterize the situation of an entity” [8], which is categorized into primary and secondary context. Primary context consists of time, location, identity and activity. Secondary context types are various and application specific. They are extracted from primary context. Making use of user context helps to find their intention of the query and as a result to disambiguate it.

Rest of the paper is organized as follows. Details of the proposed approach for query expansion are described in section 2. Section 3 illustrates the experimental results, and finally conclusion is presented in section 4.

II. THE PROPOSED QUERY EXPANSION MODEL

In this section the details of the proposed approach to query expansion is described. The model considers users’ requirements in a way that each user receives response
according do their interests. For these users context is exploited to characterize them and their interests.

Users are required to register to the system and fill out their profile at the first time they are logging in to the system. In the subsequent times, they are identified by the system. A user can change his profile information in any time he enters the system. The profile information consists of name, username, password, age and sexuality. In addition, the system stores purchase history for each user automatically.

A. Context identification

At first the context information that describe the interaction environment between the user and the system are acquired, which consists of:

- User role in the environment: User role determines main activity of a user in its current environment. For example role of a woman in the office may be a computer engineer but her role is a matron at home. In addition, her requirements are different in the office and at home. So if she searches the query “java” in the office, she probably aims “java programming language” instead of other meanings of java such as coffee or an island. Hence, user role can be exploited to disambiguate multi-meaning queries.
- User location: Local position of the user.
- User interests: User interests are entered by users to the system and kept as a vector by the system. The system augments this vector by inserting hypernyms of the words. For example if “tennis” be an interest of a user, the system inserts the concept “game” which is the hypernym of “tennis” to the user interest vector. Interest vector of a user is saved for their subsequent returns. Users can modify their interest and also profile information every time they login to the system.

B. Concept vector constitution

After logging in to the system and typing a query, query concepts, context concepts and user history vectors are constituted and each concept is given a specified weight. In the following we describe the way of constituting the concept vectors.

At first we utilize Word Net to map the user query to the concepts available in the ontology. For this, for each query word, all the synonyms, hypernyms, hyponyms, meronymy and holonymy are inserted to the Q1 concept vector, as illustrated in Fig1.

This procedure is also performed to map the mentioned context information to the concepts of the ontology. We represent the concept vector of the role context by Vec(R), the concept vector of the user location by Vec(L) and the concept vector of the user interest by Vec(I). For example considering “user role”, Vec(R) consists of synonyms, hypernyms, hyponyms, meronymy and holonymy of the user role.

Afterward, the weight of each concept is computed and assigned, as described in the next subsection.

The insertion of hypernyms and hyponyms is performed up to two levels, because expanding in more levels increases the processing time and cost.

C. Weighting

We use semantic similarity to assign weights to the concepts of the vectors. Semantic similarity is a criterion for indicating the closeness of two concepts in an ontology. Eq1 shows the formula that we have exploited to compute semantic similarity between concepts c1 and c2 [9]

\[
pl(c_1, c_2) = \frac{1}{\text{length}(c_1, c_2)}
\]

, where \(\text{length}(c_1, c_2)\) indicates length of the shortest path between \(c_1\) and \(c_2\) in the hierarchy.

In constructing query concept vector (Q1) and other vectors, for each keyword, semantic related words are inserted with their corresponding weights to the concept vectors. The weight is equal to the semantic similarity of the keyword and the concept that is to be included to the vector. This procedure is accomplished for other vectors too. Afterward, in all the vectors, concepts are sorted according to the similarity degree.

We utilize a different approach for weighting the concepts of the history and interest vectors, because the concepts of these vectors may have been located in completely different semantic domains. For example interest vector of a user may be composed of “book, tennis, music, and computer game”. Since these words, are used in different contexts, we propose an approach to find the correct context. For this, if user query has just one word, the semantic similarity of each of the concepts of the vector with this word is computed and is regarded as the weight of the concept, else if user query has multi words, the semantic similarity of each of the concepts of these vector with words of query are computed.

![Fig. 1 Concept vector constitution](image-url)
D. Selecting candidate words

After constituting weighted vectors in the previous stage, we want to find appropriate words to expand the query. For this, each word of the Q1 is compared with the words of other vectors. If a word \( W_i \) is included in the Q1 and is also existed in one or more of the other vectors, all the weights of it are added and the resulted sum is regarded as the new weight of \( W_i \). Subsequently, \( W_i \) with the new weight is placed in the final vector Q. As a result, iterated words, which are considered more important, gain more weights. In the situation that a word such as Q1 does not exist in any other vector, its weight remains unchanged and is assigned to this word in the final vector Q. At last, Q is sorted according to the descending order of weights and the words with highest weights are selected for query expansion.

For better explanation, consider a user which has entered the keyword “ball” as a query and his interests have been declared as “book, tennis, music, computer game”. Therefore, the Q1 vector which consists of synonyms, hypernyms, hyponyms, meronymy and holonymy will include more than 10 words. In this vector the words are weighted and sorted by means of similarity criterion. Vec(I) consists of user interests and especially the concept “tennis”. Consequently after query expansion, the system suggests the expression “tennis ball” to the user.

In the case that Q1 does not have any intersection with any of the other vectors, the query is augmented by the words of purchase history vector of the user. This is because usually the current purchase of a user has relation with their previous ones.

The system in default shows 10 expanded results, however this quantity is changeable.

III. EXPERIMENTAL DATA AND RESULTS

In this section we consider experimental results. For the evaluation of the approach we have chosen to use the Lucene search engine as a baseline. Therefore, this search engine has been implemented and the capability of expanding the query based on the described approach has been added to it.

The dataset clueweb09 from TREC\(^1\) which consists of about 1 billion web pages crawled in January and February 2009 has been selected. This dataset covers web content in 10 languages such as English, Chinese, Spanish and etc. 20 queries and their results on this dataset is adopted from TREC site. To evaluate the importance of context, searching with these 20 queries on four categories of users which their information can be seen in Table 2, is performed. We compare the average precision of our model with keyword search method.

<table>
<thead>
<tr>
<th>User Role</th>
<th>User location</th>
<th>Interests</th>
<th>History of purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>Book-Music</td>
<td>Tennis</td>
<td>Mobile Samsung-Mobile Nokia6760</td>
</tr>
<tr>
<td>Mobile Seller</td>
<td>Office</td>
<td>Tennis</td>
<td>Mobile Samsung-Mobile Nokia6760</td>
</tr>
<tr>
<td>Programmer</td>
<td>IT Company</td>
<td>Computer Game-Movie - Book</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>Home</td>
<td>Mouse</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\) http://trec.nist.gov/

Fig. 2 through 5 show the comparison results of the proposed approach with the no-expansion keyword search in terms of the average precision. As is clear in the diagrams, extending by the proposed model yields to an increase in precision and hence an enhancement in the search results. In other words, exploiting semantic relationships and context increase the efficiency of the search engine. In average, the proposed query expansion model yields to 12% improvement in the precision mean.

Fig. 6 compares the precision – recall diagram of our model versus keyword search (retrieval without expansion).
As it can be seen, the proposed model significantly increases the number of correct retrieved documents and precision - recall rate.

IV. CONCLUSION

In this paper a model for query expansion by utilizing concept and context has been proposed. In this model the users are identified by their context such as interests and role, and query disambiguation is performed for multi-meaning queries. The experimental results show that by exploiting semantic relations and context, the precision of information retrieval increases.

Future direction of the current research consists of designing and implementing agents to extract context automatically and exploiting data mining methods for classifying users on the basis of purchase history and interests.

REFERENCES


