Query Expansion Using Persian Ontology Derived from Wikipedia

Mojgan Farhoodi, Maryam Mahmoudi, Ali Mohammad Zare Bidoki, Alireza Yari and Mohammad Azadnia

IT Research Faculty, Iran Telecommunication Research Center (ITRC), Tehran, Iran
Electronic and Computer Engineering Department of Yazd University

Abstract: Identifying topics and concepts associated with a set of documents is a critical task for information retrieval systems. One approach is to associate a query with a set of topics selected from a fixed ontology or vocabulary of terms. The core idea of this research is using Wikipedia articles and associated pages to make a topic ontology for this purpose. The benefit of this method is that Wikipedia is an online free-content encyclopedia which is developed through a social process and kept current by the Wikipedia community. In this paper the Persian Wikipedia has been analyzed in accordance to its articles and the category link graphs to extract a Persian pseudo-ontology. Thereafter, the created ontology has been applied through a query expansion algorithm to improve the performance of an information retrieval system. Our experiments show that it is possible to improve the precision of the information retrieval system by queries expansion based on Wikipedia.

Keywords: Ontology, query expansion, Wikipedia, information retrieval

INTRODUCTION

With the rapid development of World Wide Web (WWW), extremely huge amount of information is now online and accessible by web users. Finding the required information on Web pages became more and more important and critical. Statistically, about 80 percent of users go through search engines to find their required information in web [1]. Most Search engines operate based on keywords of a query and web pages are considered as bag-of-words without perception of page creator's purpose and its concept.

As long as users have sufficient knowledge about what they want and the related concepts, the results of search engines are appropriate. Otherwise if they don't know about correct keywords, they must spend a lot of time to find related concepts and some suitable keywords for their queries. For this purpose, some methods such as query expansion have been presented [2]. Query expansion methods have been studied for a long time and several methods, such as Ontology-based [3-5], relevant feedback [6, 7], Local co-occurrence method [3, 8] and Latent Semantic Indexing (LSI based) [9, 10] have been proposed.

One of query expansion approaches is applying ontology which semantically relates and expands the queries according to their concept. By using this approach, query expansion is done semantically and users are able to have a faster access to their required information. For this purpose, Fu, Navigli and Andreou have been presented various methods and algorithms [11-13].

Since ontology making is a time consuming and costly process and also there is no proper Persian ontology, we have proposed an automatic way using Persian Wikipedia to generate a Persian ontology. The idea to bring semantics into Wikipedia is not new; several studies on this topic have been carried out in the last few years. Wikitology is one this research using Wikipedia to make English Ontology [14]. In other work the semantic relationships between English Wikipedia categories has been applied [15].

Bonino, Baesa, Sandhya, Shamsfard and Tomassen proposed different information retrieval systems using ontology as tools to expanding the queries [16-20]. In above approaches, it is very important to consider how the proposed conceptual systems are used and combined with some tradition information retrieval systems.

To improve the performance of Persian information retrieval, this paper investigates a query expansion algorithm using Persian Wikipedia-based ontology. In the proposed algorithm, by mapping the existent concepts in the queries to the graph of the Persian pseudo-ontology, the related information has been extracted to expand the queries. The experiments show a satisfactory improvement in the performance and accuracy of Persian information retrieval system.
The rest of the paper is organized as follows. First, in section 2, we give an introduction to the Wikipedia system as a comprehensive encyclopedia. In section 3 we describe the importance of query expansion process in information retrieval. Subsequently, in section 4, 5 we discuss thoroughly the proposed algorithm for generating a Persian ontology based on Wikipedia and also query expansion process. In section 6 we show the result of experiments in this regard. Finally we conclude and outline future research direction in section 6.

ROLE OF WIKIPEDIA IN CONSTRUCTING ONTOLOGY

Basic definition: Wikipedia is a free, online multilingual encyclopedia project, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Virtually any Internet user can create or edit a Wikipedia webpage. This editing policy has a positive impact on both the quantity (fast-growing number of articles) and the quality (potential mistakes are quickly corrected within the collaborative environment) of this online resource and make wiki environment as a panel of experts, authors and reviewers [21]. Anyone who wants to make knowledge available to the public can contribute an article. Anyone who encounters an article is able to correct errors, augment its scope, or compensate for bias.

Wikipedia is an open source project is covered by GNU Free Documentation License (GFDL), a copyright license permitting the redistribution, creation of derivative works and commercial use of content while authors retain copyright of their work. This is also a benefit of Wikipedia which makes it easily available for the research.

Constructing ontology based on wikipedia: The basic entry in Wikipedia is an article (or page), which defines and describes an entity or an event and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. These articles have been organized similar to a Thesaurus structure [2]. The important aspect of Wikipedia is the categorization and linkage within its content. Pages in Wikipedia are explicitly assigned to one or more categories. Categories should present major topics and their main use within Wikipedia is in finding useful information.

The core idea is to use Wikipedia's articles and associated pages as a topic ontology. Namely, each article in Wikipedia describes a single entity and the links among the pages shows the semantic relation between various entities. In another word, the documents available in Wikipedia are organized in hierarchical categories defined by its creators. Hence, Wikipedia can be considered as a graph in which the related pages (or concepts) are linked to each other. So we utilize this characteristic to construct ontology. The benefits of the approach are that the ontology terms are developed through a social process, maintained and kept current by the Wikipedia community, represent a consensus view and have meaning that can be understood by reading the associated pages.

Our ontology is Persian and is in the computer field. The details are explained in the following sections.

Query expansion: Although adding more keywords in query will specify the user’s interests, but knowing the right and specific keywords is not a simple work. Query expansion tries to extend the coverage of the retrieval to the related documents that do not necessarily contain the same words as the query.

In fact, query expansion is the process of reformulating a requested query to improve the performance of information retrieval systems. In the context of web information retrieval, query expansion involves evaluating a user's input and expanding the search query to match additional documents. Query expansion involves techniques such as: i) Finding synonyms of words; ii) Finding all the various morphological forms of words by stemming; iii) Fixing spelling errors; iv) Re-weighting the terms in the original query.

One of the previous works is Ontology-based query expansion. Ontology is resource, which provides the relation information between two concepts. The relation types include coordination, synonyms, hyponyms and other semantic relation. Some of the former works [3, 4] show that WordNet can be used as an ontology in query expansion, but it is strongly depending on the characteristics of queries. Even for some queries, the more expansion will be resulted in a worse performance.

The method of relevant feedback modifies query terms based on the distribution of the terms in the relevant and irrelevant documents retrieved in response to the original query. This method is a powerful technique and can improve the retrieval result in most cases [6, 7]. However, this method is relying on the first-retrieved top-relevant documents. If the first result is not good, relevant feedback will result in even worse results.

Local co-occurrence is a probabilistic method, which is based on the co-occurrence frequency of the words in the training corpora. This method is proved to be useful in IR system [7, 21] but it can not deal with...
meaning clustering. For instance, ‘computer’ and ‘software’ appear together several times in the corpus and the same as ‘software’ and ‘hardware’. Because of no co-occurrence between ‘computer’ and ‘hardware’, in the method, the correlation degree between them is zero or a very small value. It’s quite unreasonable.

Latent Semantic Indexing (LSI) is a powerful method, which can be implemented by two kinds of algorithms, i.e. singular value decomposition [9] and probabilistic LSI [10]. The method builds a semantic space, map each term into this space and cluster automatically according to the meaning of terms. However, it is difficult to control the query expansion degree and the modified queries may contain lots of irrelevant terms, which can be seen as noise.

This paper is investigating an ontology-based query expansion using Wikipedia, so called Persian pseudo-ontology. As mentioned before, one of the common usages of ontology is user query expansion with lessening query ambiguity in order to provide appropriate understanding of the user’s intention for the system. In this paper, user query is enriched with ontology concepts in order to make user’s request closer to his/her priorities.

The proposed architecture is composed of two main parts:

- Persian ontology construction
- Query expansion

**PERSIAN ONTOLOGY CONSTRUCTION**

Since ontology making is a time consuming and costly process and also there is no proper Persian ontology, we have proposed an automatic way to generate Persian pseudo ontology. In the proposed method there is a mapping between the model engineering space and the ontology engineering space, namely a mapping from Persian Wikipedia to a Persian ontology. This transformation results in what we call Persian pseudo ontology, as the structure of this ontology basically resembles the original meta-model and typically does not represent concepts as explicitly as ontology engineering principles would advise to do.

Every Wikipedia page consists of the following parts and we used them for constructing our ontology:

- **Title:** it introduces the concept that the page content describes it.
- **Keywords:** Some Wikipedia pages contain a number of keywords that are related to page title.
- **Text:** the text of every page describes its title. Some of the concepts existed in the text are linked to pages with the same title. These concepts are related to the title of current page. For example if the title of one page is "chip", concepts like "capacitor", "resistor" and "transistor" appear in the text as links to pages with the same titles.
- ‘See Also’ Links: this part introduces other pages with topics related to the current page. These Anchor texts are considered relevant concepts to page title.
- **Category:** this part indicates the category(s) of the page. Every page can be classified to one or more categories.

Consequently the linkage between the pages is meaningful and we can say that the relationship between these pages actually shows their conceptual relationship; so using this characteristic, we can construct an ontology with a "Is-Related-To" relation type.

To construct the ontology, it is very important to consider a correct weighting system for different relations between pages, for example the relation of page title doesn't have the same weight as the keyword or the concept of the text has. In other word, the generated ontology is a directed graph in which the concepts are nodes and the relation of the nodes are weighted.

Figure 1 shows the generating process of the Persian ontology:

As it can be seen in the Fig. 1, for ontology generating, there are three main components as follow:

**Wikipedia parser:** in the first step, according to the predefined domain (We consider ‘Computer’ domain in this paper) of Persian Wikipedia, this component gathers the related pages and then processes the page content to extract title, keywords, existing links, ‘see also’ links and category. Finally information or extracted metadata will be saved as a text file in metadata repository.

**Metadata processor:** This component takes the text files from metadata repository one by one. Then after extracting their concepts, a unique code is assigned to
every concept. The type of concept relations will be also defined in this stage. In the proposed algorithm three kinds of relationship have been presented:

- The relationship between the title and the keywords
- The relationship between the title and concepts in text (that appear as links).
- The relationship between the title and ‘see also’ links.

All of these relations are “Is-Related-To” type. Some anchor texts in Wikipedia pages aren’t the same with the title of the linked pages. For example, there may be a link with “Computers” anchor text that links to a page with “Computer” title. In this relationship the contexts of “Computers” and “Computer” is the same. So we define a type of relationship named “Same-As” which is different from three previous relationships and it is considered in generating the proposed ontology. Using this relationship, some Persian language features such as forms of writing a concept in a singular or plural form will be also covered; so concepts like “Information Bank”, “Information Banks”, “Databases”, and “Database” are considered the same.

**Ontology generator:** This component takes the concepts and information about relationship and finally generates the ontology. There are various types of relationships; so various weights are assigned to the links.

To construct the ontology, we use a directed graph with weights in the form of a matrix in which rows and columns are codes of concepts. Namely each cell of this matrix indicates the weight of relationship between the concept number \(i\) and the concept number \(j\). Figure 2 shows an example of ontology construction base on parsing Wikipedia pages:

As it is shown in Fig.2, concepts are nodes of ontology with a unique code. On the other side, the relationship between concepts is directed and its weight is specified. For example, the weight between “Operating System” and “Memory Management” shows that the page with “Operating System” title has “Memory Management” concept in its ‘see also’ links. The weight 0.5 between these two concepts means that the page with “Operating System” title contains “Memory Management” concept in its keywords. All of these relationships are “Is-Related-To” type. The relationship between “Computers” and “Computer” is a mutual “Same-As” relationship with the weight 1.

![Fig. 2: Persian ontology sample](image)

**Fig. 2: Persian ontology sample**

![Fig. 3: Query expansion process](image)

**Fig. 3: Query expansion process**

**THE PROPOSED QUERY EXPANSION ALGORITHM**

As mentioned before, using ontology is one of the approaches to user query expansion and removing the ambiguity of the query in such a way that system can understand user's intention. In the proposed approach, the existent concepts in the query are mapped on to the ontology graph based on Wikipedia. In this way, the Wikipedia-based Persian pseudo ontology is used to expand user queries and improve the search results.

As shown in Fig. 3, a query expansion component gets the user query, processes it and then determines the existing concepts. Thereafter, the related concepts of the user query have been extracted by using the ontology graph and considering the relationship weights. Then, the query has been expanded and submitted to the search engine. Finally, the results of search engine are returned to the user.

To expand the query, we use vector method. The query vector is formed based on the existing concepts in the query and it is an array with a length equivalent to the number of existing concepts in the ontology. The index of each element in the query vector represents the corresponding concept code. If the existent concept(s) in the query is available in the ontology, the corresponding element value becomes 1 and otherwise 0.
To form the expanded query vector, the following formula is used:

\[ \tilde{q}_e = \tilde{q} + \alpha \tilde{q}O + \beta \tilde{q}O^2 + \gamma \tilde{q}O^3 + \delta \tilde{q}O^4 \]

\[ \alpha, \beta, \gamma, \delta \geq 0 \quad \alpha + \beta + \gamma + \delta = 1 \]

\( \tilde{q} \): Query vector; \( \tilde{q}_e \): Expanded query vector; \( O \): Ontology matrix.

And we have:

\[ O_{i,j} = \begin{cases} \text{1} & \text{The weight of each concept with itself and similar concepts} \\ \text{0} & \text{If two concepts don't have any relations} \end{cases} \]

\[ W_t \] The weight between the title and keywords;

\[ W_e \] The weight between the existing concepts in the text;

\[ W_s \] The weight between the title and the existing concepts in the 'see also' links.

\[ qO \] indicates that a concept points directly to which concepts in the graph. In other words, considering the ontology graph, \( qO \) finds out-links of the given concept. \( \tilde{q}O^2 \) is able to detect the concepts which are a level further away from the given concept. For example, if in the ontology graph "Database" concept points to "Relational Model" concept and "Relational Model" concept points to "Table" concept, \( \tilde{q}O^2 \) enables the system to detect the relation between "Database" and "Table". Therefore, for each concept two levels of ontology can be examined, similarly the in-links of a concept can be considered. \( \tilde{q}q \) and \( \tilde{q}O \) elements are also used in the formula 1. The difference between \( \tilde{q} \) and \( \tilde{q}q \) is that \( \tilde{q} \) shows the concepts which are linked by the given concept and \( \tilde{q}q \) shows the concepts which links to the given concept. The result of considering these issues in the query expansion is that if node number \( i \) links to the node number \( j \) and vice versa, then these two concepts are related to each other very much and have higher weight. In Fig. 2 For instance, "Operating System" and "Memory Management" concepts have this situation.

Due to the relevancy rate between two concepts which is depending on their distance in graph and type of the links (i.e. in-links or out-links), \( \alpha, \beta, \gamma \) and \( \delta \) coefficients are defined to show their importance level.

The expanded query may contain many concepts with different weights; therefore, to get better results, a threshold value must be used and concepts with lower weights are omitted from query expansion process.

Figure 4a and 4b show the ontology matrix and its square respectively. Rows and columns of ontology matrix represent concept codes and their corresponding value show the weight between them. It is obvious that each concept in relation to itself has the weight 1.

Now, we suppose that user enters "Operating System" term as input query. After receiving the user query and processing it, system forms the query vector. The query vector is formed as follows: if the existing concept(s) in query is available in the ontology, its corresponding element is considered as 1 and otherwise as 0. In the above example, the query vector is formed as follows:

\[ \tilde{q} = [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0] \]

The third index of the element is 1 which means that there is a concept with code 3 in the user query.
Table 1: Test results

<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>Precision</th>
<th>Expanded query</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fuzzy logic</td>
<td>0.6</td>
<td>Artificial intelligence, expert systems, fuzzy logic, logic, fuzzy sets</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>information theory</td>
<td>0.3</td>
<td>Information theory, information, philosophy of information</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>knowledge engineering</td>
<td>0.6</td>
<td>Knowledge engineering, artificial intelligence, semantic web, knowledge management</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>web</td>
<td>0.2</td>
<td>Web, web standards, internet</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>network cart</td>
<td>0.6</td>
<td>Routing, network cart</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>integrated circuit</td>
<td>0.5</td>
<td>Integrated circuit, chip, resistance, capacitor, semi conductor</td>
<td>0.4</td>
</tr>
<tr>
<td>7</td>
<td>assembly</td>
<td>0.6</td>
<td>Assembly, programming, reverse engineering</td>
<td>0.7</td>
</tr>
<tr>
<td>8</td>
<td>LAN</td>
<td>0.8</td>
<td>LAN, network, Ethernet</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>oracle</td>
<td>0.3</td>
<td>Oracle, database</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>genetic algorithm</td>
<td>0.4</td>
<td>Genetic algorithm, computer science, inheritance, Darwin</td>
<td>0.7</td>
</tr>
<tr>
<td>11</td>
<td>spyware</td>
<td>0.5</td>
<td>Spyware, computer, internet</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>CPU</td>
<td>0.6</td>
<td>CPU, processor, computer, microprocessor</td>
<td>0.7</td>
</tr>
<tr>
<td>13</td>
<td>genum</td>
<td>0.3</td>
<td>Genum, operating system, open source, linux, BSD</td>
<td>0.6</td>
</tr>
<tr>
<td>14</td>
<td>firewall</td>
<td>0.6</td>
<td>Firewall, hardware, software, network</td>
<td>0.6</td>
</tr>
<tr>
<td>15</td>
<td>internet explorer</td>
<td>0.4</td>
<td>Internet explorer, web explorer, Microsoft, windows</td>
<td>0.7</td>
</tr>
</tbody>
</table>

The expanded query vector is formed using the formula (1) with coefficients of 0.9, 0.1, 0 and 0 respectively:

\[ \tilde{q}_e = [0 \ 0 \ 2.03 \ 0.55 \ 0.07 \ 0.77 \ 0 \ 0] \]

Considering the expanded query vector, it can be concluded that the term such as "Operating System", "Software", "Computer" and "Memory Management" concepts are related to the user query. If a threshold has been set as 0.7, the expanded query contains "Operating System" and "Memory Management" concepts.

**EXPERIMENTAL RESULTS**

To test the query expansion algorithm together with Persian pseudo ontology, following experiments have been done. After expansion for various queries, they were submitted to Google search engine. To evaluate the performance of system, the 20 first returned results in Google were evaluated by expert computer users. The relevancy rate of a page is considered as 1 and 0 which respectively represents a relevant or irrelevant page.

Table 1 shows the test sample queries along with the expanded queries and precisions before and after expansion. Precision in Table 1 is defined as the number of relevant documents in the top 20 results retrieved by Google.

The precision of the query expansion algorithm is evaluated by 100 queries in their top 20 ranked documents before and after using the proposed algorithm. Figure 5 shows the result of experiments before and after query expansion. The experimental results show that the precision of retrieval system is improved by utilizing the proposed query expansion.

![Fig. 5: Comparison between precision before and after query expansion](image)

We give different values to \( \alpha, \beta, \gamma, \delta \) and test precision of results. Table 2 shows the obtained precision for 100 queries in respect to their values. According to the result of experiments which has been reflected partially in Table 2, we found that the following values are the best in the query expansion process:

\[ \alpha = 0.7, \beta = 0.2, \gamma = 0.05, \delta = 0.05, \text{Threshold}>0.7 \]

Similarity, after doing various tests, we found that the following weights for concepts make the result of retrieval better:

Table 2: Values of coefficients

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
<th>( \delta )</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.10</td>
<td>0.30</td>
<td>0.1</td>
<td>0.35</td>
</tr>
<tr>
<td>0.6</td>
<td>0.20</td>
<td>0.10</td>
<td>0.1</td>
<td>0.42</td>
</tr>
<tr>
<td>0.7</td>
<td>0.20</td>
<td>0.05</td>
<td>0.05</td>
<td>0.53</td>
</tr>
<tr>
<td>0.9</td>
<td>0.05</td>
<td>0.05</td>
<td>0.0</td>
<td>0.51</td>
</tr>
</tbody>
</table>
• Weight of the relation between title and keywords: 0.6
• Weight of the relation between title and the concepts in the text: 0.5
• Weight of the relation between title and the concepts in ‘see also’ links: 0.7

As it is demonstrated in Table 1, most of the results have higher precision when the query expansion is enriched. But in some queries, extended query cause to decrease the expected precision.

Since the presented ontology is constructed with the link amongst Wikipedia pages, the content quality and existing links in pages have direct impact on generating better ontology.

CONCLUSION AND FUTURE WORK

After an overview on approaches using ontology to improve information retrieval systems, an approach for query expansion using a Wikipedia-based Persian ontology has been proposed. The benefit of this approach is that the automatically generated ontology is related to a live, online and growing encyclopedia.

To construct the ontology, the concepts have been considered as nodes and links between different nodes define the relations between them. To make the relations meaningful, a weighting system is applied. In the weighting system, title and keywords, title and the text concepts, title and ‘see also’ links concepts, have respectively the weight of 0.6, 0.5 and 0.7.

The vector space method is used for query expansion, in which elements is assigned 0 or 1 with regards to the existence or non-existence of user query concepts in the ontology. Since the concepts of user query may not be available in the set of ontology concepts, it is suggested to use individual words instead of concepts in user query.

The proposed query expansion method together with the Wikipedia-based Persian ontology is applied in Google search engine to show the improvement of Persian searching result. 100 queries in their top 20 ranked documents are evaluated before and after using the proposed approach. The results of experiments show that the precision of the retrieval system is improved by utilizing the proposed query expansion method.

When the concepts in query have not been related to the concepts in ontology, the possibility of extending the query doesn’t exist. It is better to separate the query into its components and calculating their similarity with the concepts of the ontology. In the future works, the process of extending the query according to the similarity ratio of existing words in the query and the available concepts in ontology will be considered.

REFERENCES

5. Rila Mandala, Takenobu Tokuanga and Hozumi Tanaka, 1999. Combining multiple evidence from different types of thesaurus for query expansion. SIGIR.