A New Method for Applicant of Explicit Semantic Analysis and Word Sense Disambiguation in Concept-based Information Retrieval

Rezvan mohamad rezaei#1, mehran mohsenzadeh#2, mashallah abbasi dezfooli#3

#1 Department Of Computer Engineering Science And Research Branch Islamic Azad University Khouzestan Iran, r.mohamadrezai@khouzestan.srbiau.ac.ir.

#2 Department Of Computer Engineering Science And Research Branch Islamic Azad University Tehran Iran, mohsenzadeh@srbiau.ac.ir.

#3 Department Of Computer Engineering Science And Research Branch Islamic Azad University Iran, Abbasi_masha@yahoo.com.

Abstract

Previous Information retrieval (IR) systems based on keywords to retrieve and index documents. They may return inaccurate results when different keywords are employed to illustrate the same concept in the documents and in the queries presented by Users. In Concept-based retrieval methods have tried to tackle these troubles by using concept-based comparison between documents and queries. Therefore, accurate concept extraction of documents and queries improves performance of IR systems. In this research, we introduce a new concept-based query semantic analysis approach based on Wikipedia-based Explicit Semantic Analysis. We propose that first specify the given context of query words by using Wikipedia-based concept network named wikinet to query words sense disambiguation. Then rely on given context create related concepts of query that they will be compare to concepts of documents. Because the main aim of this paper is to provide a correct interpretation and semantic relatedness analysis of query words, we use of correlation of computed relatedness scores with human judgments. Evaluation shows that the proposed method provides improvements compared to the existing semantic analysis methods.

Keywords- concept-based information retrieval, query, query sense disambiguation, Wikipedia, wikinet.

1.1 Introduction

In the present age, which is called the information age, there is a lot of information available for users in the web. Extension of electronic documents in the web despite their all advantages, create some problems such as finding intended information in the extreme and varied information. Therefore, some methods are required to organize and search for information. One of the methods for finding desired information is information retrieval. Information retrieval (IR) system is finding documents that satisfy information need (query) of user from within large collection of documents [1]. IR systems intend to provide the maximum related documents to query of user. primal IR systems were used by retrieval experts, therefore initial IR method was rely on keywords manually assigned to documents, and on difficult Boolean queries. After 1970 automatic indexing and natural
language queries obtained popularity and non-skilled users more widely used of these IR systems [2]. This model were called keyword-based information retrieval model. In this model, keyword lists are used to illustrate contents of information objects. Documents were indexed by automatically considering all terms in them as independent keywords, in what is known as the Bag-of-Words (BOW) representation, and query formatting was simplified to a short natural language formulation. Keywords list is a description that does not say anything about semantic relationships between keywords [3]. A serious weakness of such systems is that they can be misguide by the ambiguity of terms (i.e. polysemy) and ignore relationships among terms (e.g. synonym) [4]. IR researchers attempted to resolve the synonymy problem by expanding the original query with synonyms of query keywords [5]. However, the relationship between the keywords chosen by the users and those used by the authors often extends beyond simple synonymy. So a pertinent document may contain related information to user query but it does not mention of any directed synonym of any of the query keywords. To handle synonym problem, new query expansion methods that rely on corpus-based evidence were suggested. Such methods showed significant improvement, but require manual tuning in order not to adversely affect performance: too few expansion terms may have no impact, and too many will cause a query drift [2]. To tackle polysemy problem, the main proposed method was to apply automatic word sense disambiguation algorithms to documents and queries. Disambiguation methods use resources such as the Wordnet or co-occurrence data to find the possible senses of a word and map word occurrences to the correct sense. These disambiguated senses are then used in indexing and in query processing, so that only documents that match the correct sense are retrieved. The inaccuracy of automatic disambiguation is the main obstacle in achieving significant improvement using these methods [2]. A new generation IR model is Concept-based information retrieval that purposes to handle problems of keyword-based IR systems. Generally, a content of an information object is described by a set of concepts in this model. Concepts can be extracted from the text by categorization. Crucial in this model is existence of a conceptual structure for mapping descriptions of information objects to concepts used in a query [3]. So a concept-based IR system performs retrieval in that concept space. Quality of query interpretation and the extracted concepts of query in retrieval process of concept-based IR system are very important. Users enter a query that it is understandable to a human being. Unfortunately, in a large number of cases such queries are not handled well by the IR systems. Therefore this problem creates difference between what users enter their queries and the desired documents that they want.

In this paper, we focus on query processing and propose a new method for accurate query interpretation and representation because the enhancing quality of query interpretation directly affects the retrieved results [6]. If a query is interpreted correctly then retrieved documents will be more relevant to query and performance of concept-based IR systems will improve. Therefore, we introduce a new concept-based query semantic analysis approach based on Wikipedia-based Explicit Semantic Analysis. The Explicit Semantic Analysis (ESA) [7] is a promising approach for explicit semantic analysis. We have used the semantic relatedness of query words to their sense disambiguation. This approach firstly specifies the given context of query words with the aim of query sense disambiguation. The given context finds using the concepts and their relations are extracted from Wikinet. Wikinet is a multi-lingual concept network obtained automatically by mining for concepts and relations and exploiting a variety of sources of knowledge from Wikipedia [12]. Then based on specified context of query, it interprets each word of query as weighted vector of Wikipedia based concepts. Our main goal in this
paper is that query interpretation algorithm of concept-based IR systems to Human mental algorithm are close.

The remainder of this paper is organized as follows. Section 2 firstly provides background on ESA and section 3 introduces the wikinet. Sections 4 and 5 describe the details of the proposed algorithm and the evaluation results. Finally, Section 6 concludes the paper.

2.1 Explicit Semantic Analysis

Explicit Semantic Analysis, or ESA [7], is a recently proposed method for computing semantic relatedness. ESA represents meaning in a high-dimensional space of concepts automatically derived from large-scale human-built repositories such as Wikipedia. ESA is based on the assumption that in Wikipedia an article corresponds to a semantically distinct concept. Each article is pre-processed by tokenization, stemming and stop word removal and is represented as a vector of words that occurs in this article weighted by their TFIDF score. These weights quantify the strength of association between words and concepts.

Firstly, a text fragment is represented as a vector using TFIDF scheme, then semantic interpreter that implements as a centroid based classifier [11] ranks all the Wikipedia articles by their relevance to the text. To speed up semantic interpretation, ESA uses an inverted index, which maps each word into a list of concepts in which it appears. The semantic interpreter iterates over the text words, retrieves corresponding entries from the inverted index, and merges them into a weighted vector of concepts that represents the given text. Entries of this vector reflect the relevance of the corresponding articles to text. This method compares weighted vectors of the Wikipedia articles related to a particular term or portion of text using the cosine metric to compute semantic relatedness. Gabrilovich and Markovitch show that ESA outperforms other existing approaches. Compared with the previous state of the art, using ESA results in notable improvements in correlation of computed relatedness scores with human judgments. However, ESA doesn’t specify a given semantic context of the words and uses the similar vectors for ambiguous word.

Consider, for example (jaguar, cat) and (jaguar, car). The semantic context of (jaguar, cat) is animal and the semantic context of (jaguar, car) is automobile. However, ESA to semantic relatedness analysis builds similar vectors for “jaguar” without paying any attention to the semantic context of words [10]. In this paper, we proposed that before use of explicit semantic analysis for query words interpretation, we initially specify the given context of query words using the Wikinet.

3.1 Wikinet

In recent years, researchers have realized the huge potential of Wikipedia as a source of semi-structured knowledge and several systems have used it as their main source of knowledge. Wikipedia provides a massive and relatively high-quality collection of text and encyclopedic knowledge. The use of a knowledge repository as large and diverse as Wikipedia creates a powerful concept network, well suited for semantic analysis. First, Wikipedia's broad coverage of a huge range of topics, and second, mapping from a massive aggregation of natural language terms to the concepts in which they occur, produce a powerful classifier to automatically map any text fragment to this concept network. Finally, the use of a Wikipedia generates meaningful and human readable concepts that can provide additional reasoning for the researcher and for system users.

Several approaches have been used to extract semantic information from Wikipedia.
Wikinet [12] is a very large scale, multilingual concept network, obtained by exploiting several facets of Wikipedia. The resource consists of a language independent concept base extracted from Wikipedia articles and categories, and the relations between them. Relations between concepts are extracted from the several sources of knowledge from Wikipedia some explicitly (articles, categories and their links, infoboxes), some implicitly (category names). This coverage of multiple pieces of information differentiates it from similar endeavors and resources extracted from Wikipedia. WikiNet is supposed to be used to complement WordNet with knowledge about numerous named entities (which were outside the scope of WordNet) as well as general concepts and numerous relations. In this paper, we made use of wikinet for specifying given semantic context of query words because Wikipedia is only partly structured and cannot be used by a computer without some processing. This concept network provides concepts and relations that related to any word of query. Then we compare concepts and their relations and generate a given context of query.

4.1 Proposed Method

The main aim of our proposed method in this paper is query words sense disambiguation and correct query interpretation by using Wikipedia-based explicit semantic analysis. As described above, ESA has got a great success in the semantic relatedness analysis based on Wikipedia but ESA neglects the context of words, thus it cannot exactly determine the desired sense of an ambiguous word. Therefore, we introduce our method that first specifies the given context of words query semantically, and then extracts corresponding concepts for each Word under the given context. At result the retrieved documents will be more relevant to query and the performance of concept-based IR systems will improve. We use the wikinet for Specifying given context of words. This method proposed for short query because shorter queries are more pervasive than longer ones in the web domain and the average query length is around 2.3 words [6]. The detail of the method is as follows. To illustrate the proposed method, we assume that the query presented as (w1, w2).

4.1.1. Step1: Query Words Category Specification

The Initial step of our method is query words Category Specification. The ‘QueryWordsCategorySpecification’ algorithm chooses a given category for related concepts by extracting concepts and their relations from WikiNet. The algorithm firstly finds indexes of words in query from index.wiki file and saves in Indexw1, Indexw2. If a word of query has several indexes, it is an Ambiguous word [12], therefore we must specify the intended sense of the word by the user. Then it extracts relations of indexes from data.wiki file and each concept of Indexw1 and Indexw2 compared based on their IS-A relations. If concept1 and concept2 have IS-A relation with a given concept, they are related concepts and have relation on a given context [12], [17], [18]. Although the IS-A relation to a common concept does not exist between some of concepts, they may be related. Thus, after considering above condition, we extract from inlinks.wiki file the list of concepts that link to concept1 and concept2 through their corresponding pages and save in InLinkconcept1 and InLinkconcept2. Now it compares the IS-A relations of InLinkconcept1 to IS-A relations of the concept2 and vice versa. The given concepts that have IS-A relation to both words indicates the common context for query words [17], [18].
Finally, we choose a given category for related concepts that the category's articles have relation with concepts query words. We define $C_1$ and $C_2$ as the set of categories assigned to concept $1$ and concept $2$, respectively. We then determine the semantic relatedness value for each category pair $(c_k, c_l)$ with $c_k \in C_1$ and $c_l \in C_2$. Then we uses the notion of a lowest common subsumer of two nodes lcs $(c_k, c_l)$ i.e. the minimum for path based [13]. The given category is a common semantic environment for query words.

Algorithm : QueryWordsCategorySpecification $(w_1, w_2)$

For each index $w_{1i} \in \text{Index}_{w1} = \{ \text{index}_{w11}, \ldots, \text{index}_{w1n} \}$ do
  For each index $w_{2j} \in \text{Index}_{w2} = \{ \text{index}_{w21}, \ldots, \text{index}_{w2m} \}$ do
    Let $\text{Is-alIndex}_{w_{1i}}$ be a set of concepts that have is-a relation with index $w_{1i}$
    Let $\text{Is-alIndex}_{w_{2j}}$ be a set of concepts that have is-a relation with index $w_{2j}$
    CommonConcept$_{\text{indexw1i, indexw2j}} \leftarrow \text{Is-alIndex}_{w_{1i}} \cap \text{Is-alIndex}_{w_{2j}}$
    If CommonConcept$_{\text{indexw1i, indexw2j}} = \emptyset$ then
      Let $\text{InLink}_1$ be a set of concepts that link to index $w_{1i}$
      Let $\text{InLink}_2$ be a set of concepts that link to index $w_{2j}$
      For each $\text{InLink}_{1x} \in \text{InLink}_1 = \{ \text{InLink}_{11}, \ldots, \text{InLink}_{1s} \}$
        Let $\text{Is-alIndex}_{\text{InLink}_{1x}}$ be a set of concepts that have is-a relation with $\text{InLink}_{1x}$
        CommonConcept$_{\text{indexw11, indexw2j}} \leftarrow \text{Is-alIndex}_{\text{InLink}_{1x}} \cap \text{Is-alIndex}_{w_{2j}}$
      For each $\text{InLink}_{2y} \in \text{InLink}_2 = \{ \text{InLink}_{21}, \ldots, \text{InLink}_{2t} \}$
        Let $\text{Is-alIndex}_{\text{InLink}_{2y}}$ be a set of concepts that have is-a relation with $\text{InLink}_{2y}$
        CommonConcept$_{\text{indexw11, indexw2j}} \leftarrow \text{Is-alIndex}_{\text{InLink}_{2y}} \cap \text{Is-alIndex}_{w_{1i}}$
      If CommonConcept$_{\text{indexw11, indexw2j}} \neq \emptyset$ then
        GivenCategory $\leftarrow$ CommonCategory(categories of Index$_{w11}$, Index$_{w2j}$)
        CommonConcept $\leftarrow$ CommonConcept$_{\text{indexw11, indexw2j}}$
      If CommonConcept $= \emptyset$ then
        GivenCategory $\leftarrow$ CommonCategory(categories of Index$_{w1}$, Index$_{w2}$)
    Return GivenCategory

4.1.2. Step2: Query Words Context Word Generation

The ‘QueryWordsContextWordGeneration’ algorithm specifies words of query context based on the given category. In this algorithm, we measure the cosine metric of the TFIDF vectors of the query words and the articles of given category to determine their relatedness. Finally, in order to utilize the context information in the ContextRelatedArticle, we select a set of B words with the highest TFIDF weight in ContextRelatedArticle to represent the context of the query words. Similar to [10], In proposed method, we use standard attribute selection techniques such as information gain to identify words that are most characteristic of a concept versus of all other concepts.
Algorithm: QueryWordsContextWordGeneration \((w_1, w_2)\)

\[
\text{GivenCategory} \leftarrow \text{QueryWordsCategorySpecification} \,(w_1, w_2)
\]
\[
\text{Text}(w) \leftarrow \text{Combine} \,(\text{Text}(w_1), \text{Text}(w_2))
\]
\[
\text{Vector}(w) \leftarrow \text{TFIDF} \,(\text{Text}(w))
\]

\[\text{For each concept}_i \in \text{GivenCategory} = \{\text{concept}_1, \ldots, \text{concept}_n\} \text{ do}\]
\[
\text{Text} \,(\text{concept}_i) \leftarrow \text{AttributeSelection} \,(\text{Text} \,(\text{concept}_i))
\]
\[
\text{Vector} \,(\text{concept}_i) \leftarrow \text{TFIDF} \,(\text{Text} \,(\text{concept}_i))
\]
\[
\text{Value} \,(\text{concept}_i) \leftarrow \text{DistanceMetric} \,(\text{Vector} \,(\text{concept}_i), \text{Vector}(w))
\]

\[\text{Let ContextRelatedArticle be a set of A articles that have highest value} \,(\text{concept}_i)\]

\[\text{For each concept}_i \in \text{ContextRelatedArticle do}\]
\[
\text{Text} \,(\text{ContextRelatedArticle}) \leftarrow \text{add} \,(\text{Text} \,(\text{concept}_i))
\]
\[
\text{Vector} \,(\text{ContextRelatedArticle}) \leftarrow \text{TFIDF} \,(\text{Text} \,(\text{ContextRelatedArticle}))
\]

\[\text{Let ContextRelatedtext be a set of B words that have highest TFIDF in the Text} \,(\text{ContextRelatedArticle})\]

\[\text{Return ContextRelatedtext}\]

4.1.3. Step 3: Query Word Concepts Generation

After acquiring the ContextRelatedtext by Step 2, we must generate the corresponding concepts for each word of query with the help of the ContextRelatedtext that generated semantic context of the query and GivenCategory. The ‘QueryWordConceptsGeneration’ algorithm generates the corresponding concepts for each word of query.

Algorithm: QueryWordConceptsGeneration

\[
\text{ContextRelatedtext} \leftarrow \text{QueryWordsContextWordGeneration} \,(w_1, w_2)
\]
\[
\text{Text}(w_1) \leftarrow \text{combine} \,(\text{Text}(w_1), \text{ContextRelatedtext})
\]
\[
\text{Text}(w_2) \leftarrow \text{combine} \,(\text{Text}(w_2), \text{ContextRelatedtext})
\]
\[
\text{Vector}(w_1) \leftarrow \text{TFIDF} \,(\text{Text}(w_1))
\]
\[
\text{Vector}(w_2) \leftarrow \text{TFIDF} \,(\text{Text}(w_2))
\]

\[\text{For each concept}_i \in \text{GivenCategory} = \{\text{concept}_1, \ldots, \text{concept}_n\} \text{ do}\]
\[
\text{Text} \,(\text{concept}_i) \leftarrow \text{AttributeSelection} \,(\text{Text} \,(\text{concept}_i))
\]
\[
\text{Vector} \,(\text{concept}_i) \leftarrow \text{TFIDF} \,(\text{Text} \,(\text{concept}_i))
\]
\[
\text{Value}_{w_1} \,(\text{concept}_i) \leftarrow \text{DistanceMetric} \,(\text{Vector} \,(\text{concept}_i), \text{Vector}(w_1))
\]
\[
\text{Value}_{w_2} \,(\text{concept}_i) \leftarrow \text{DistanceMetric} \,(\text{Vector} \,(\text{concept}_i), \text{Vector}(w_2))
\]

\[\text{Let Vector}_{w_1} = \{\text{Value}_{w_1} \,(\text{concept}_1), \ldots, \text{Value}_{w_1} \,(\text{concept}_n)\}\]

\[\text{Let Vector}_{w_2} = \{\text{Value}_{w_2} \,(\text{concept}_1), \ldots, \text{Value}_{w_2} \,(\text{concept}_n)\}\]

\[\text{Return Vector}_{w_1}, \text{Vector}_{w_2}\]

after the interpretation query, concept-based IR systems can retrieve documents that their extracted concepts are same with extracted concepts of query words. Whatever the query interpretation be more correct, the results of concept-based IR systems are more accurate. To illustrate process of this method, we show the common concepts and the concepts of ambiguous word vector for two sample queries: (jaguar, tiger) and (jaguar, car). The common concepts that are extracted from IS-A relations of (jaguar, tiger) are animal, mammal, fauna, biota And Big cat. And the extracted common concepts of (jaguar, car) are Car manufacturers, Automotive companies, Vehicle, Road transport, Motor vehicle manufacturers. These queries contain ambiguous word “jaguar”. In our method, the semantic environment of query specified by information of a common context therefore it
is capable of performing word sense disambiguation. Table 1 contains the ten highest-valuing Wikipedia concepts in Vectorw1 for "jaguar". If concepts in the vector are sorted in the decreasing order of their value, the top ten concepts are the most relevant ones for the word of query. Figure 1 makes clear our approach and illustrates the process of interpretation of query by using wikinet and Wikipedia.

Table 1
The Ten Highest-Valuing Wikipedia Concepts In Vectorw1 For "Jaguar"

<table>
<thead>
<tr>
<th>(jaguar, tiger)</th>
<th>(jaguar, car)</th>
</tr>
</thead>
<tbody>
<tr>
<td>jaguar</td>
<td>Jaguar car</td>
</tr>
<tr>
<td>cougar</td>
<td>Jaguar S-Type</td>
</tr>
<tr>
<td>leopard</td>
<td>Jaguar E-Type</td>
</tr>
<tr>
<td>genus Panthera</td>
<td>Jaguar X-Type</td>
</tr>
<tr>
<td>Big cat</td>
<td>Jaguar C-Type</td>
</tr>
<tr>
<td>Snow leopard</td>
<td>Jaguar XK</td>
</tr>
<tr>
<td>Panther hybrid</td>
<td>Jaguar XJ</td>
</tr>
<tr>
<td>tiger</td>
<td>Jaguar XF</td>
</tr>
<tr>
<td>puma</td>
<td>V8 (V8 engine)</td>
</tr>
<tr>
<td>European lion</td>
<td>Luxury vehicle</td>
</tr>
</tbody>
</table>

Fig. 1 the process of proposed method
5.1. Empirical Evaluation

We implemented our approach for to evaluate the results of query interpretation using this method in concept-based IR systems. Because our goal in this paper is to improve concept-based IR systems performance based on correct query interpretation and the proposed method performs query sense disambiguation and query interpretation rely on semantic relatedness analysis, so we use the datasets and measures that relevant to semantic relatedness analysis methods. Finally, the results of the proposed method are compared with the results of existing semantic analysis methods. For implementation of our approach, we use a Wikipedia snapshot as of "January 15, 2011" (http://dumps.wikimedia.org/enwiki/20110115). Also we use the "20110115" version of WikiNet (http://www.hits.org/english/research/nlp/download/wikinet). With WikiNet's 3.7 Million concepts and 40 Million relations (instantiating 656 relation types), efficiency in data management becomes an issue. Manual analysis of the data is also problematic. WikiNetTK [15] addresses both these issues. A fast data management is the basis for an easy-to-use visualization component. For using wikipedia first, we preprocess the Wikipedia dump with stemming, stop word removal, frequency of words calculation and attribute vector discovering. So Wikiprep parses the Wikipedia XML dump (http://www.cs.technion.ac.il/gabr/resources/code/wikiprep). Upon removing small and overly specific concept (those having fewer than 100 words and fewer than 5 incoming or outgoing links), In order to speed up the experiment process, We used the Natural Language Toolkit to parse each article as single words and their frequency of seeming in this page. We processed the text of these articles by first tokenizing it, removing stop words and rare words (occurring in fewer than 3 articles), and stemmed the remaining words, this giving up distinct terms, which were used to represent article as attribute vectors. Additionally, because the dataset is large enough to be divided into training set and testing set, the pattern selection for optional parameter estimation was executed as a grid search through cross-validation on the training data. in implementation, the value of A is 9 and the value of B is 16 [10].

5.1.1. Dataset And Evaluation Procedure

To evaluate the accuracy of query interpretation method, the results of method should be comparing with human judgement. To compute semantic relatedness of query words, we compare their final vectors using the cosine metric i.e. SR=Cosine (Vector_{w1}, Vector_{w2}). Whatever the results of method is closer to human judgments, so the semantic relatedness between words have been identified with great accuracy and the method is More efficient. Evaluating word relatedness is a natural ability humans have and is, therefore, considered a common baseline. To assess word relatedness, we use the WordSimilarity-353 benchmark dataset, available online [14], which contains 353 word pairs. Each pair was judged, on average, by 13-16 human annotators. This dataset, to the best of our knowledge, is the largest publicly available collection of this kind. In this paper we consider each word-pair as a query. Spearman rank-order correlation coefficient between the computed relatedness scores by experiment and the corresponding human judgments on WordSimilarity-353 benchmark were used to compare computed relatedness scores with human judgments.

5.1.2. Result

We compare the results of proposed method with achieved results of other semantic relatedness analysis method to demonstrate the efficiency of the proposed method. Whatever the result of a method is closer to 1 indicating that this method can detect better
the semantic relatedness between words query and perform query words sense disambiguation. Table 2 shows the result of applying our method for making judgment relatedness of word-pairs (that are considered as queries) of WordSimilarity-353.

### Table 2
**Result On Correlation With Human Judgement**

<table>
<thead>
<tr>
<th>algorithm</th>
<th>Correlation with humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet [16]</td>
<td>0.33-0.35</td>
</tr>
<tr>
<td>Roget's Thesaurus [16]</td>
<td>0.55</td>
</tr>
<tr>
<td>LSA [19]</td>
<td>0.56</td>
</tr>
<tr>
<td>WikiRelate! [9]</td>
<td>0.19-0.48</td>
</tr>
<tr>
<td>WLVM [8]</td>
<td>0.45</td>
</tr>
<tr>
<td>ESA-wikipedia [7]</td>
<td>0.75</td>
</tr>
<tr>
<td>SAESA-wikipedia [10]</td>
<td>0.81</td>
</tr>
<tr>
<td>Our method</td>
<td>0.81</td>
</tr>
</tbody>
</table>

The results of experiment show that our method makes substantial improvements over prior studies. Our method, compared to statistically methods such as LSA, which only uses statistical co-occurrence information from a large unlabeled corpus of text, uses the knowledge resource that is collected and organized by humans. Also, compared to the methods relying on lexical resources such as WordNet and Roget's Thesaurus [16] that cover only a small fragment of the language lexicon, our method leverages knowledge resources that are orders of magnitude larger and more comprehensive. Our proposed approach improved also Wikipedia-based approach such as WikiRelated! [9] and WLVM [8]. That is because our approach represents intended sense of each word as a weighted vector of Wikipedia concepts, and semantic relatedness is then computed by comparing the two concept vectors. Besides, the obtained results of ESA [7] method is most close to our method and shows that our method is more effective to computing semantic relatedness. Our method also achieves similar results to that of the SAESA [10]. However, SAESA in order to generate given context of words considers all concepts in Wikipedia that it needs for a very long time. But our method finds the given context of query words just by searching the information of wikinet, and based on the context, it expresses the intended meaning for the ambiguous word and compute semantic relatedness. However, Empirical evaluation confirms that using our method leads to substantial improvements in query interpretation in concept-based IR systems.

### 6. Conclusion
In this paper, we proposed an improved approach to query interpretation in concept-based IR systems based on explicit semantic relatedness analysis of query words by using the concepts and relations of Wikipedia. Our approach first by using of a Wikipedia-based concept network named wikinet, generates the given context of query words and specifies the intended sense of ambiguous word. Then corresponding concepts of query are extracted based on given context of query words. The Results show that the semantic relatedness of query words are identified correctly by proposed method. Therefore after query interpretation, concept-based IR system will retrieve documents that have same context with query and contain extracted concepts of query. So performance of concept-based IR systems will improve.
References


