TOWARD A BETTER SEARCH RESULT: A STUDY OF THE REFINEMENT OF USERS' QUERIES IN INFORAMTION RETRIEVAL

長崎大学大学院工学研究科
YUZANA WIN

Nowadays, it is an important but complex task to get valuable information by searching the Web. With the rapid increase of information, users often perceive the difficulty of accessing the rich information resource effectively and of obtaining the information associated with their needs accurately. When users want to find the information relevant to their needs, the search results may not be relevant due to the inability of the queries to represent the needs accurately. In this thesis, we propose a method for supporting query refinement by using the users' queries of search result, i.e., query words closely related to a specific topic, extracted from a retrieved set of documents. Our method aims to explore useful information or knowledge relevant to users' needs and interests from real-world data sources, e.g. newspaper or website articles, research papers, blog entries and patent applications, and thus to provide a useful system for information retrieval. In order to improve search effectiveness, our method apply a technique called "query expansion" that can be used for obtaining additional terms relevant to a given search keywords. It is usually used to help information searchers express their intentions more accurately and increase the precision of search results. Therefore, we propose a method for improving search results by query expansion. In this thesis, our method consists of three ways that can be used for query expansion.

The first way, we propose a method for extracting important phrases as word *n*-grams from research paper titles. The extracted phrases are expected to be fruitful in query expansion for academic information retrieval. In extracting a phrase, here, we introduce a new measure of phrase importance called *technical phrase frames*. The outcome of our method can provide *phrase frames*, not phrases in the ordinary sense. Phrase frames are phrases with wildcards (*) that may be substituted for any word. By substituting the wildcards of a phrase frame with words, we can obtain a complete phrases. *Firstly*, we extract word trigrams (a sequence of three words) from research paper titles. *Secondly*, we construct a co-occurrence graph of the extracted trigrams. To construct the co-occurrence graph, we consider the direction of edges in two ways: *forward* and *reverse*. In the forward and reverse co-occurrence graphs, the trigrams point to other trigrams appearing after and before them in a paper title, respectively. *Thirdly*, we assign weights to the edges of the co-occurrence graphs based on Jaccard similarity between trigrams. *Fourthly*, we apply the weighted version of PageRank

algorithm. Consequently, we can obtain the top-ranked trigrams associated with the higher PageRank scores. Many of the top-ranked trigrams given from these two co-occurrence graphs can be regarded as important phrases. Our method is special in the following sense. The method extracts two types of phrases, each of which realizes a different query expansion, i.e., the expansion to the left and the expansion to the right. For example, our method gives "a framework for" and "in sensor networks" as its outcome. The phrase "a framework for" can expand queries like "clustering", "classification", etc., to the left and give more specific queries like "a framework for clustering", "a framework for classification", etc. The phrase "in sensor networks" can expand queries like "clustering", "classification", etc., to the right and give more specific queries like "clustering in sensor networks", "classification in sensor networks", etc. After that, we evaluate the extraction of phrases as additional features in the paper title classification task. We use SVM (Support Vector Machine) for classification and check whether the extraction of phrases can improve the classification accuracy. The experimental results show that our method works well. Finally, we describe a search system for research paper titles that embodies the proposed method to show the expanding queries with the extracted phrases which can be used for query expansion. As a result, by expanding queries with the extracted phrases, users can get search results relating to specific topics.

The second way, we propose a method for non-trivial compound word as new words extraction from Myanmar text. Our main purpose is to find compound words that can be added into the existing Myanmar dictionary. The outcome of our method are new compound words which are not described in the Myanmar dictionary. When we use existing search engines, we enter only a few words to form a query. If the user needs compound word as a query, the search system automatically recognizes the query word as new words, not just the sequence of several words. In order to obtain the compound words, our method consists of two steps. In the first step, we extract maximal substrings from Myanmar news articles. Maximal substrings are defined as the substrings whose number of occurrences are reduced by any of its extensions. As the second step, we make a post-processing of maximal substrings, because the resulting maximal substrings contain noisy characters. In our post-processing, we reduce the number of maximal substrings and remove maximal substrings whose prefixes and suffixes are meaningless characters. We keep only the substrings that consist of words from the existing dictionary. As a result, we obtain the substrings as candidates of new compound words that can be added into the existing Myanmar dictionary.

After that, we examine the candidate compound words extracted by our method as additional document features in an unsupervised manner. Therefore, we apply K-means clustering for obtaining document clusters and check whether the compound words extracted by our method can improve the clustering results. We compare the compound words given by

our method with the two other methods. The experimental results show that the quality of clustering results was improved. Finally, we implement a prototype of a user interface for news articles that embodies the proposed method to show many new words appear as compound words which can be used for query expansion. A search system this type of query expansion can give search results relating to specific topics.

The third way, we propose a method for proper names extraction from Myanmar text. Our method aims to extract proper names that provide important information on the contents of Myanmar text. For example, the query "Myanmar politics" is expanded by the proper names like "Daw Aung San Suu Kyi" and "Min Ko Naing". We can obtain more specific queries like "Myanmar politics Daw Aung San Suu Kyi" and "Myanmar politics Min Ko Naing". If the additional terms are relevant to the query concept, the precision of search results are improved. Therefore, proper names extraction is an important research area of query expansion. Our method consists of two steps. In the first step, we extract topic words from Myanmar news articles by using latent Dirichlet allocation (LDA). In the second step, we make a post-processing, because the resulting topic words contain some noisy words. Our post-processing, first of all, eliminates the topic words whose prefixes are Myanmar digits and suffixes are noun and verb particles. We then remove the duplicate words and discard the topic words that are contained in the existing dictionary. Consequently, we obtain the words as candidate of proper names, namely *personal names*, *geographical names*, *unique object names*, *organization names*, *single event names, and so on.*

We evaluate the extracted proper names as additional document features in K-means clustering. The evaluation is performed both from the subjective and quantitative perspectives. We compare the precision, recall and F-score of proper names extracted by our method with those extracted by latent semantic indexing (LSI) and rule-based method from the subjective perspective. It is shown that our method is better remarkable improvement for the extraction of proper names than LSI and rule-based method. From the quantitative perspective, the experimental results show that the document clusters given by our method are better than those given by LSI and rule-based method in precision, recall and F-score.

Based on our study and experiments, we understand that search results can be improved in our three cases by using additional terms. Moreover, our method can extract expansion terms without using additional data such as WordNet. The improved search results can give users a great help in Web searching and Web usage. The results show that our method is useful in discovering the documents consisting of specific and precise topics. As a result of the experiments, we believe that the important phrases, compound words and proper names provided by our method are good candidates for query expansion in information retrieval task.