Distance Based Similarity Thesaurus Construction for Arabic Document Retrieval

Ahmad Mustofa 51162010

Myrna Ermawati 5116201022

Addien Haniefardi 51162010

----------------------------------------------------------------------------------------------------------------------

Abstract

In this paper, we propose a new approach to the similarity thesaurus construction method used for query expansion in document retrieval. The method was used for constructing an Arabic similarity thesaurus. Regarding the terms co-occurrence, we calculate the average of minimum distance between two terms co-occur in a same document. Our experiment use this similarity thesaurus to determine top-n similar terms for query expansion.

1. Introduction

In order to perform document retrieval, an important issue is how to improve the performance of retrieval systems. One of such techniques is query modification. There are two main classes of query modification methods. The first class is query expansion which the system reformulates the user query [1, 2] by adding extra terms and re-weighting the query terms. The second class concentrates only on re-weighting the query terms.

Other concern in document retrieval is how to represent each document into document vectors from terms frequently occurring in the documents. The system will continue finding the similar document terms from the query given and retrieve top documents contain most similar terms. However, it ignores the conceptual similarity between terms. Ambiguity in the meaning of terms (for example, the word “saving”, and ‘‘fund” may have the same meaning) can prevent the system from retrieving more relevant document. Thesaurus is mostly used to address this problem.

A thesaurus may consists of a set of weighted term associations which are based on the hypothesis that terms have relationship if they co-occur often in the documents [3]. These associated terms are then used to query expansion. Many researchers also discussed topics about terms association in behavior of their co-occurrence in documents. Terms that often co-occur close to each other are most likely to be related than if they just co-occur in the same document, or within some distance of each other [11]. Two terms that co-occur in a short document have bigger similarity value than ones that co-occur in long document [12]. Lexical co-occurrence means the co-occurrence of two terms when appearing within some distance of each other.

We consider the possible chance of this co-occurrence in a long document is greater than the possibility in a short document. By this consideration, we proposed a new similarity measurement for thesaurus generation that involve the distance between 2 terms, and the document’s length. In this paper, we propose a method for constructing an automatic thesaurus for Arabic document by calculating and involving the distance between document term and query term. The similarity measure will compute the distance between terms. The n-most similar terms will be used to construct automatic thesaurus.

The rest of this paper is organized as follows: Section 2 provides an overview of the related studies and researches. Section 3 presents our approach to automatic thesaurus construction and the usage for query expansion in detail. Section 4 reports our experimental results. Finally, Section 5 concludes the paper.

1. Related Studies
   1. Thesaurus Construction Based on Co-occurrence

Researchers have proposed a number of approaches that deal with the problem of vocabulary mismatch. One of the main approaches is using a thesaurus to expand user query using synonymous words in order to retrieve more relevant items. Qiu & Frei [5] had an experimental study if the term's indexing weight might be used in measuring two terms' similarity like similarity thesaurus. According to their study, the retrieval effectiveness would be better, because the collocation in a short document plays a more important role than that in a long document.

Many researches focus on co-occurrence analysis [6] that is based on the idea that similar terms co-occur frequently in the document collection. Peat and Willett [7] concluded that co-occurrence analysis should be used only for the identification of terms that co-occur less frequently in the collection. Chen and Lynch [8] had an experiment on applying an algorithmic approach to the automatic generation of thesauruses based on statistical correlation analysis.

After that, Tseng [9 and 10] developed a similar approach for automatic thesaurus construction from Chinese documents based on co-occurrence analysis with modified Dice coefficient. The author found that long documents tend to yield weaker Dice coefficient in comparison with short documents. Also, Tseng argued that terms co-occur in the same documents may have no relationships if they are apart from each other. This is the reason why his approach limits the terms to be related to the terms that are occurring in the same sentence.

Schutze and Pedersen [11] presented a method for automatically constructing a thesaurus based on lexical co-occurrence. Lexical co-occurrence means the co-occurrence of two terms when appearing within some distance of each other. The authors argued that terms that often co-occur close to each other are most likely to be related than if they just co-occur in the same document. Al Qabbane [12] stated that two terms that co-occur in a short document have bigger similarity than ones that co-occur in long document.

* 1. Similarity Thesaurus Construction

In the similarity thesaurus construction approach, each term is represented by a vector in the document vector space, . The weight represents the relation between the term and the document. Qiu and Frei [5] adopted the well-known normalized () schema (1) as in the Vector Space Model [4] and modified it to reflect the indexing approach. The weight is computed using Equation (2)

(1)

(2)

Where, n is the number of documents and is the occurrence frequency of term in document while is the maximum frequency of the term in all documents. is the inverse term frequency for document and calculated using (3).

(3)

* 1. Query Expansion and Document Ranking

The basic idea of similarity thesaurus is to calculate the similarities between two terms on the basis of their co-occurrence in a document corpus. This approach is based on the association hypothesis that related terms tend to co-occur in documents in the corpus. As a result, this type of automatic thesaurus consists of a set of weighted term associations. The similarity thesaurus is constructed by determining the similarities of all the term pairs (ti ,tj ).

Using similarity thesaurus, a query is then expanded by looking for terms that have high similarity with the entire query terms. All the terms pairs calculated in the thesaurus construction will be ranked in descending order according to their sim(ti, tj). The top N terms are used as expansion terms.

1. Proposed Approach
   1. Average Distance Between Terms

In the previous research, Tseng’s approach to terms similarity that limit the co-occur terms in only one sentence obtain a better result in the field of thesaurus generation [9, 10]. Pederson [11] and Al Qabbane [12] has stated that two terms co-occur in a short document have bigger similarity than ones that co-occur in long document. Learning these studies, we proposed a method of thesaurus generation approach that involve the distance between terms and document’s length.

Regarding the terms co-occurrence, we calculate the average of minimum distance between two terms co-occur in a same document. Let pos-i = [] and pos-j = [] are list of position where query-term and term occur in document-k, where R is frequency of query-term and T is frequency of term in a document. Total minimum distance of those two terms then is written as equation (3).

(3)

(4)

* 1. Automatic Similarity Thesaurus Construction

Given n is the number of documents in the corpus. is term weight of in document k which value is calculated using well-known normalized tf-idf as equation (1). By involving the average of minimum distance as written in (4) in our approach, the similarity between two terms then can be calculated by using (5). The smaller the average distance between two terms, the more similar they are.

(5)

The is the inverse term frequency for document-k written in equation (3). We adopt the inverse term frequency in this approach to indicates that a document length also have a contribution since we involve average distance between two terms co-occur in the same document. If two terms co-occur in a long document, the probability that two terms co-occur in the same document is greater than their co-occurrence in a short document.

1. Experiment and Result
   1. Arabic Document Collection

For our experiment, document in Arabic is taken from a collection of books that appear on Maktabah Syamilah software. Maktabah Syamilah software has widely used in Islamic boarding school (pesantren) as media learning for the students. Simply, this software is a digital version of yellow books that are widely taught in schools.

As an input data, Arabic document corpus used in our system consists of more than 6000 documents. The contents of the books saved in our system are assumed to be a one-page document. A page of a book in Maktabah Syamilah is saved as one document in our system. We could also add new document from customized page contents from one book, or *kitab*, in Maktabah Syamilah.

|  |  |
| --- | --- |
| Number of Documents |  |
| Number of Terms |  |
| Number of Unique Term |  |
| Average Number of Terms in Documents |  |

Tabel-. Characteristic of Document Corpus

* 1. Query Expansion and Document Ranking

1. Conclusion

Refferences

Additional notes \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[2] A. Ibrahim El-Khair, "Arabic Information Retrieval," Information Science and Technology, vol. 41, no. 1, pp. 505-533, 2007.

[5] L.S. Leah and B. Lisa, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis," Special Interest Group on Information Retrieval, pp. 275- 282, 2002

Keywords: Automatic Thesaurus, Query Expansion, Arabic Document Retrieval, Information Retrieval, Optimized Term Document Similarity