Shahjalal University of Science and Technology, Sylhet

Department of Computer Science and Engineering

Query Expansion For Bangla Search Engine Pipilika

A.F.M.Kamruzzaman
Reg. No.: 2014331029
4^{th} year, 2^{nd} Semester

Md. Rezaul Islam
Reg. No.: 2014331044
4^{th} year, 2^{nd} Semester

Department of Computer Science and Engineering

Supervised by:

Dr. Farida Chowdhury
Associate Professor
Department of Computer Science and Engineering

14th February, 2019

Query Expansion For Bangla Search Engine Pipilika

A Thesis Submitted to the

Department of Computing Science and Engineering

Shahjalal University of Science and Technology

Sylhet - 3114, Bangladesh

in partial fulfillment of the requirements for the degree of

B.Sc.(Engg.) in Computer Science and Engineering

By

A.F.M.Kamruzzaman

Reg. No.: 2014331029

4^{th} year, 2^{nd} Semester

Md. Rezaul Islam

Reg. No.: 2014331044

4^{th} year, 2^{nd} Semester

Department of Computer Science and Engineering

Supervised by:

Dr. Farida Chowdhury

Associate Professor

Department of Computer Science and Engineering

14th February, 2019

Recommendation Letter from Supervisor

The thesis

entitled "Query Expansion For Bangla Search Engine Pipilika"

submitted by the students

1. A.F.M.Kamruzzaman
2. Md. Rezaul Islam

is a record of research work and then development of an interface carried out under my supervision.

I, hereby, agree that the thesis can be submitted for examination.

Signature of the Supervisor:

Name of the Supervisor: Dr. Farida Chowdhury

Date: 14th February, 2019

Certificate of Acceptance of the Thesis/Project

The thesis
entitled " Query Expansion For Bangla Search Engine Pipilika"
submitted by the students

1. A.F.M.Kamruzzaman
2. Md. Rezaul Islam

on 14th February, 2019
as part of the requirements of the course CSE-452, is being approved by the Department of Computer Science and Engineering as a partial fulfillment of the B.Sc.(Engg.) degree of the above students.

Head of the Dept.	Chairman, Exam. Committee	Supervisor
Dr Mohammed Jahirul Islam	Dr. Farida Chowdhury	Dr. Farida Chowdhury
Professor	Associate Professor	Associate Professor
Department of Computer	Department of Computer	Department of Computer
Science and Engineering	Science and Engineering	Science and Engineering

Abstract

Query expansion has been adopted in search engine to remove the ambiguity of queries. It aims to extend the user query with related terms to improve the search results of a search engine and the relevance of the information which are targeted to retrieve. So, Query Expansion is a method to select additional words which are related to the search keyword to make the search more accurate and efficient.

We select the additional terms after ranking the documents using BM25 document ranking method. We generate similar words for additional and search key words using a word embedding method and selected according to their semantic similarity to the original word.

Keywords: Query Expansion, BM25 Document Ranking Algorithm, Word2Vec CBOW Method.

Acknowledgements

We would like to thank the Department of Computer Science and Engineering, Shahjalal University of Science and Technology for supporting in this research. We are very thankful to our honorable supervisor Dr. Farida Chowdhury and Pipilika R&D Team for supporting us.

Dedication

We would like to dedicate our research to our parents and to our younger brothers.

Contents

Abstract I

Acknowledgement II

Dedication III

Table of Contents IV

List of Figures VI

1 Introduction **1**

1.1 Background 1

1.2 Reasons to Motivation in Query Expansion 2

1.3 Research Questions 2

2 Query Expansion **3**

2.1 What Is Query Expansion? 3

2.2 Why need Query Expansion? 3

2.3 Types of Query Expansion 4

3 Background Study **5**

3.1 BM25 Algorithm : 5

3.2 Word2Vec CBOW Model : 6

3.3 Related works : 8

4 Methodology **9**

9

4.2 Flow Diagram Of Full Procedure 12

5 Result and Analysis **13**

5.1 Result 13

5.1.1 Collected Bangla Stop Word List: 13

5.1.2 Filtered Search Key: 13

5.1.3 Retrieved Documents After Filtering: 14

5.1.4 Top 25 Documents After Applying BM25 Algorithm: 14

5.1.5 Collected Additional Terms From Top Documents: 15

5.1.6 Final Output: 15

5.2 Analysis 16

6 Conclusion **17**

6.1 Discussion 17

List of Figures

3.1 The architecture of Word2Vec method 7

4.1 Flow Diagram of Query Expansion Method 12

5.1 Stop Word 13
5.2 INPUT 13
5.3 Filtered Documents 14
5.4 Ranked Documents 14
5.5 Additional Terms 15
5.6 Expanded Query 15

Chapter 1

Introduction

1.1 Background

Information retrieval (IR) aims to provide a user easy access in interested information. From a collection of documents, term-based document retrieval methods generate queries that retrieves user's interested documents. An Information retrieval system ranks the retrieval documents. It returns relevant documents to the user. This method faces some problems like that too short queries problem, lexical gap problem or or too ambiguous queries etc. For this, our Pipilika search engine cannot fulfill user interests. To solve these problems, we propose a query expansion method to capture more semantic information about user's interests.

It aims to select new expanded words to a query to improve the performance of an IR system. To improve the performance of IR systems, a Query expansion (QE) techniques must be needed for Pipilika Search Engine.
However, in the query expansion process two important questions need to be answered:

1. How should the additional words be selected?

2. How should the similar words be selected for additional and search key words?

For this, firstly we retrieved documents from the corpus which has included search key terms. Secondly, we have ranking these documents using BM25 term-based document ranking method. Thirdly, we extract additional terms from top ranked documents. Fourthly, we select similar words for additional and search key words using Word2Vec CBOW model.

As Pipilika Search Engine is the first Bengali search engine in Bangladesh, it is highly expected that the search engine should provide an efficient searching service to its user. A better Query expansion method can provide a better Searching experience.

1.2 Reasons to Motivation in Query Expansion

We are motivated in Query expansion for Pipilika Search Engine because of two reasons. Those are listed below-

1. **Unexpected Query Results: :** Usually, we cannot get the exact result which we expect in Pipilika Search Engine when we use a portion of our information as a search key.
   So, we observed the need of a better Query expansion technique to be applied here.
2. **Lack of Intelligent Query Results:** So far we have observed that the query results we retrieve by performing a search are not intelligent search results. Those results are unable to fulfill users expected search information in some cases where the complex search keys are used to search in the search engine. So, we are motivated to apply intelligent techniques to do better search operations in the search engine.

1.3 Research Questions

Here are some questions that we are investigating in our research:

1. What are the limitations of the current information retrieval process?
2. How does improved Query Expansion impact in performing efficient search operation?
3. What are the techniques to be applied for the Query Expansion in this search engine?
4. How should the documents be ranked?
5. How should the additional words be selected?
6. How should the similar words be selected for additional and search key words?
7. What are the improvements by applying proposed Query Expansion techniques in this research?

Chapter 2

Query Expansion

2.1 What Is Query Expansion?

Query Expansion(QE) is to extend the user query with similar terms to improve search engine result for better user experience. It aims to predict most suitable similar words to add with the query to increase information retrieval effectiveness.
In search engines, query expansion is applied for expanding the search query for matching with additional documents. [1].

Query expansion(QE) works by adding terms to the user's original query to improve the information retrieval effectiveness. For adding terms in the original query, it is done by including similar words which are used in the relevant documents.

Query expansion is a methodology which is applied in the computer science field particularly within the information retrieval and natural language processing etc.

2.2 Why need Query Expansion?

A major problem in Information Retrieval(IR) is the mismatch between query terms and relevant documents terms which satisfy the user needs information. Query expansion (QE) is a popular technique which is used to remove this vocabulary mismatch.

There are some other pitfalls can be observed in traditional query result processing system

[1]https://en.wikipedia.org/wiki/Query_expansion

where efficient Query Expansion process is not applied.

2.3 Types of Query Expansion

There are basically two types of Query Expansion:

1. **Global methods :** Global methods is one of the first process to produce effective progress through query expansion method. Work independently of the query and result set. Find word relationships in the corpus. Use an external source (spelling, thesaurus etc). One of the earliest global methods is term clustering , which groups document terms into clusters based on their co-occurrences. Global methods needs corpus-wide statistics such as statistics of co-occurrences of pairs of terms[1].

2. **Local methods :** Analyze the documents returned by the base query.

Chapter 3

Background Study

Before we jump into discuss our own method we should take a look at BM25, Word2Vec CBOW method.

3.1 BM25 Algorithm :

Document queries have been produced using a BM25 method. Each words BM25 score is computed according to a set of documents that the user is interested. The term weights are estimated by

$$W(t) = \sum_{i}^{n} IDF(q_i) * \frac{f(q_i, D) * (k1 + 1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})} \tag{3.1}$$

Here, we can see a few common components like q_i, IDF(q_i), f(q_i,D), k1, b, and something about field lengths. Each of these are described below:

1. q_i is the i^{th} query term.

2. IDF(q_i) is the inverse document frequency of the i^{th} query term.

$$IDF(q_i) = \ln\left(1 + \frac{(docCount + f(q_i) + 0.5)}{f(q_i) + 0.5}\right) \tag{3.2}$$

Where docCount is the total number of documents and f(q_i) is the number of documents

which hold the i^{th} query term.

3. Field length is divided by average field length in the denominator as **fieldLen/avgFieldLen**. If a document is greater than average field then the average field length gets bigger for decreasing the score and if it is shorter than average field then the average field length gets smaller for increasing the score.

4. A variable b is multiplied by the ratio of the field length. If b is bigger, the effects of the document length compared to the average length are more raised. But, if we set the value of b is 0 then the effect of the length ratio will be completely nullified and the document length would be no bearing on the score. **By default, the value of b is 0.75 in Elastic search.**

5. Finally, two components of the score which show up in the equation are k1 and $f(q_i , D)$. These components are discussed in below:

   (a) $f(q_i,D)$ is "how many times does the i^{th} query term occur in document D".

   (b) k1 is a variable which find the saturation characteristics of term frequency.

3.2 Word2Vec CBOW Model :

Word embedding is the best popular representation of document words. It is powerful to capture semantic and syntactic similarity or relation of a word with other words.
To train word embedding using shallow neural network Word2Vec is the most popular method. It was developed by Tomas Mikolov in 2013 at Google[1]. We chose the Word2Vec method for our query expansion process since it has been shown to be effective for training word embedding in huge scale text data. Word2Vec generates a vector space and projects every word to a point in that space. Similar words for a word are selected by computing its cosine distance from the original word in the vector space. For example,

$$v(king) - v(queen) \approx v(man) - v(woman) \tag{3.3}$$

[1]https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

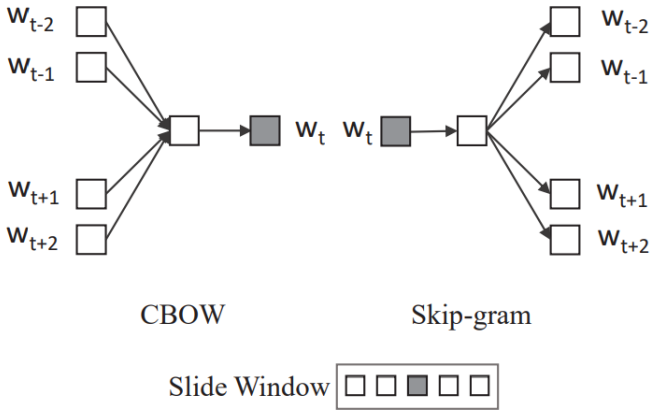

Figure 3.1: The architecture of Word2Vec method

Figure 3.1 shows that Word2Vec's architecture contains two models: a continuous bag-of-word model(CBOW) and a Skip-gram model.

1. **CBOW model :**The CBOW model try to predict the target word using the contextual words in the sliding window. Let, there have given a word sequence

$$D = \{w_{i-k}, ..., w_{i-1}, w_i, w_{i+1}, ..., w_{i+k}\} \tag{3.4}$$

where w_i is the target word, the objective of CBOW is to maximize the average log probability

$$L(D) = \frac{1}{T}\sum_{i=1}^{T} \log P_r(w_i|w_{i-k}, ..., w_{i-1}, w_{i+1}, ..., w_{i+k}) \tag{3.5}$$

where, T is the size of corpus and k is the context size of target word which implys that window size is (2k + 1). CBOW formulates the probability

$$P_r(w_i|w_{i-k}, ..., w_{i-1}, w_{i+1}, ..., w_{i+k}) = \frac{\exp(x_i.x_c)}{\sum_{w \in W} \exp(x_i.x_c)} \tag{3.6}$$

where W is the vocabulary, x_i is the vector representation of the target word, w_i and x_c is the average vector of all contextual words.

2. **Skip-gram model :** Skip-gram model aims to predicts context words using the target word in the sliding window. However, the Skip-gram model purpose is to maximize the average

log probability

$$L(D) = \frac{1}{T}\sum_{i=1}^{T} \sum_{-k \le c \le k, c \neq 0} \log P_r(w_{i+c}|w_i) \tag{3.7}$$

where k is the context size of the target word and the probability $P_r(w_{i+c} \mid w_i)$ is formulated with soft-max function which is denoted as

$$P_r(w_{i+c}|w_i) = \frac{\exp(x_{i+c}.x_i)}{\sum_{w \in W} \exp(x.x_i)} \tag{3.8}$$

where W represents the vocabulary, x_i is the vector representation of the target word w_i, and x_{i+c} is the vector representation of context word.

Both have own advantages and disadvantages. Skip Gram is good for small volume of data. On the other hand, CBOW is faster and better for large amount of data. So, we chose CBOW model for query expansion.

3.3 Related works :

Representing words using fixed Length vectors is an important step for processing texts in document retrieval process. This method is preferred for its simplicity and effictiveness. Since this method does not consider semantic information, it suffers from various problem like that data sparsity problem, the curse of dimensions problem and the lexical gap problem which makes information retrieval process difficult. Distributed word representation known as word embedding has been solved these problems. In this method, words are described as dense, low-dimensional, real-valued vectors. All dimension described the semantic and syntactic features of words. Recently, neural network (NN) algorithms has been used for learning word representations([17],[18],[15]). Several papers have focused on query expansion using word embedding. Fernando Diaz et al.[19] proposed the use of term-relatedness information generated by word embedding method. Kuzi et al.[20] proposed a query expansion method based on CBOW.

Chapter 4

Methodology

In this chapter we shall discuss about the proposed query expansion method for document ranking in detail. We have divided the method into multiple steps. Now we will each step in below:

4.1 Steps of Query Expansion[1]

1. **Removing Stop words From Search Key :**

   (a) **What are Stop words?**

   A stop word is a commonly used word that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

   (b) **Why need Stop words remove?**

   These words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to be stop words.

For this, first we generate a stop words list for Bangla.

$$StopWord = Stopword\_Retrieve()$$

[1]https://github.com/ZARIFREZAUL/QUERY-EXPANSION/

After giving input a search key in Pipilika search engine text box by a user, we remove stop words from search key.

$$INPUT = Removing\_Stopword(SearchKey, StopWord)$$

2. **Retrieving Documents from Entire Corpus :**
   We retrieved all documents from entire corpus which documents has included search key words.
   $$Document = RETRIEVING\_DOCUMENT(INPUT)$$

3. **Filtering Retrieved Documents :**
   After retrieving documents from corpus we need to filtering the documents. First of all, we need to remove unnecessary characters from the documents. Second, we need to remove stop words from the documents. Third, we need to remove duplicate documents from the document list. Fourth, removing single word documents from the document list.
   $$Filtered\_Document = FILTERING\_DOCUMENT(Document)$$

4. **Ranking Filtered Documents using BM25 Algorithms :** After filtering documents we need to rank documents using BM25 algorithm compared to INPUT.
   $$Ranked\_Document = BM25(Filtered\_Document)$$

5. **Selecting Additional Words From Top Ranked Documents :** From top ranked documents we collect additional word from the documents by removing search keyword from the documents.
   $$Additional\_Term = Term\_Selection(Ranked\_Document)$$

6. **Fetching similar words from Word2Vec CBOW model :**

   First, We load word2vec CBOW model.

$$model = Word2Vec.load(”word2vec\_model.model)$$

   Second, we collect similar words for each search key and additional term from model by computing its cosine distance from the original word in the vector space. After collecting similar words we remove those words which are spelling mistake of original words.

$$Similar\_Word = Retrieve\_Similar\_Word(word, model)$$

7. **Generating Output According To Expression :**

   The generated expanded Query results for a search key will look like below:

$$Q(Output) = (\quad S(Word2Vec(similar\_words)) \quad OR \quad Add(Word2Vec(similar\_words)) \quad)$$

   Here ,

   (a) **Word2Vec(similar_words)=Association of all similar words for a word using ’OR’**

   (b) **S(Word2Vec(similar_words))=Association of Word2Vec(similar_words) for each search key word using ’AND’**

   (c) **Add(Word2Vec(similar_words))= Association of Word2Vec(similar_words) for each additional words fetched from corpus using ’OR’**

4.2 Flow Diagram Of Full Procedure

4.2. FLOW DIAGRAM OF FULL PROCEDURE

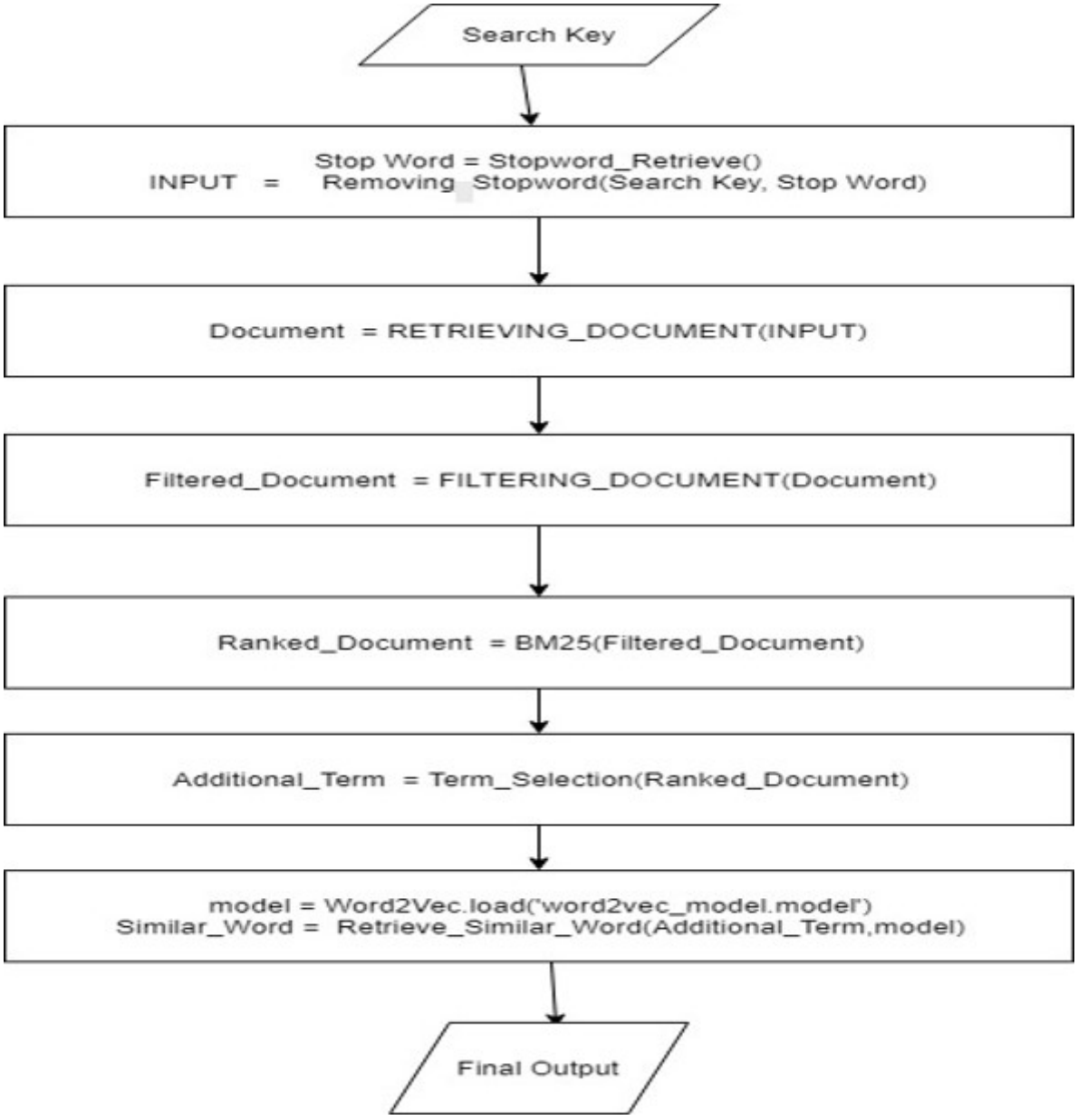

Figure 4.1: Flow Diagram of Query Expansion Method

Chapter 5

Result and Analysis

5.1 Result

5.1.1 Collected Bangla Stop Word List:

(অতএব, অথচ, অথবা, অনুযায়ী, অনেক, অনেকে, অনেকেই, অন্তত, অবধি, অবশ্য, অর্থাৎ, অন্য, অনুযায়ী, অর্ধভাগে, আগামী, আগে, আগেই, আছে, আজ, আদ্যভাগে, আপনার, আপনি, আবার, আমরা, আমাকে, আমাদের, আমার, আমি, আর, আরও, ইত্যাদি, ইহা, উচিত, উনি, উপর, উপরে, উত্তর, এ, এঁদের, এঁরা, এই, এক, একই, একজন, একটা, একটি, একবার, একে, এখন, এখনও, এখানে, এখানেই, এটা, এসো, এটাই, এটি, এত, এতটাই, এতে, এদের, এবং, এবার, এমন, এমনি, এমনকি, এর, এরা, এলো, এস, এসে, ঐ, ও, ওঁদের, ওঁর, ওঁরা, ওই, ওকে, ওখানে, ওদের, ওর, ওরা, কখনও, কত, কখা, কবে, কয়েক, কয়েকটি, করছে, করছেন, করতে, করবে, করবেন, করলে, কয়েক, কয়েকটি, করিয়ে, করিয়া, করায়, করলেন, করা, করাই, করায়, করার, করি, করিতে, করিয়া, করিয়ে, করে, করেই, করেছিলেন, করেছে, করেছেন, করেন, কাউকে, কাছ, কাছে, কাজ, কাজে, কারও, কারণ, কি, কিংবা, কিছু, কিছুই, কিন্তু, কী, কে, কেউ, কেউই, কেন, কোন, কোনও, কোনো, কেমনে, কোটি, ক্ষেত্রে, খুব, গিয়ে, গিয়েছে, গুলি, গেছে, গেল, গেলে, গোটা, গিয়ে, গিয়েছে, চলে, চান, চায়, চেয়ে, চায়, চেয়ে, চার, চালু, চেষ্টা, ছাড়া, ছাড়াও, ছিল, ছিলেন, ছাড়া, ছাড়াও, জন, জনকে, জনের, জন্য, জন্যে, জানতে, জানা, জানানো, জানায়, জানিয়ে, জানিয়েছে, জানায়, জ্যানিয়ে, জানিয়েছে, টি, ঠিক, তখন, তত, তথা, তবু, তবে, তা, তাঁকে, তাঁদের, তাঁর, তাঁরা, তাঁহারা, তাই, তাও, তাকে, তাতে, তাদের, তার, তারপর, তারা, তারই, তাহলে, তাহা, তাহাতে, তাহার, তিনই, তিনি, তিনিও, তুমি, তুলে, তেমন, তো, তোমার, তুই, তোরা, তোর, তোমাদের, তোদের, থাকবে, থাকবেন, থাকা, থাকায়, থাকে, থাকেন, থেকে, থেকেই, থেকেও, থাকায়, দিকে, দিতে, দিয়ে, দিয়েছে, দিয়েছেন, দিলেন, দিয়ে, দু, দুটি, দুটো, দেওয়া, দেওয়ার, দেখতে, দেখা, দেখে, দেন, দেয়, দেশের, দ্বারা, দিয়েছে, দিয়েছেন, দেয়, দেওয়া, দেওয়ার, দিন, দুই, ধরা, ধরে, নয়, না, নাই, নাকি, নাগাদ, নানা, নিজে, নিজেই, নিজেদের, নিজের, নিতে, নিয়ে, নিয়ে, নেই, নেওয়া, নেওয়ার, নয়, নতুন, পক্ষে, পর, পরে, পরেই, পরেও, পর্যন্ত, পাওয়া, পারি, পারে, পারেন, পেয়ে, প্রতি, প্রভৃতি, প্রায়, পাওয়া, পেয়ে, প্রায়, পাঁচ, প্রথম, প্রাথমিক, ফলে, ফিরে, ফের, বছর, বদলে, বরং, বলতে, বলল, বললেন, বলা, বলে, বলেছেন, বলেন, বসে, বহু, বা, বাদে, বার, বিনা, বিভিন্ন, বিশেষ, বিষয়টি, বেশ, ব্যবহার, ব্যাপারে, বক্তব্য, বন, বেশি, ভাবে, ভাবেই, মত, মতো, মতোই, মধ্যভাগে, মধ্যে, মধ্যেই, মধ্যেও, মনে, মাত্র, মাধ্যমে, মানুষ, মানুষের, মোট, মোটেই, মোদের, মোর, যখন, যত, যতটা, যথেষ্ট, যদি, যদিও, যা, যাঁর, যাঁরা, যাওয়া, যাওয়ার, যাকে, যাচ্ছে, যাতে, যাদের, যান, যাবে, যায়, যার, যারা, যায়, যিনি, যে, যেখানে, যেতে, যেন, যেমন, রকম, রয়েছে, রাখা, রেখে, রয়েছে, লক্ষ, শুধু, শুরু, সাধারণ, সামনে, সঙ্গে, সঙ্গেও, সব, সবার, সমস্ত, সম্প্রতি, সময়, সহ, সহিত, সাথে, সুতরাং, সে, সেই, সেখান, সেখানে, সেটা, সেটাই, সেটাও, সেটি, স্পষ্ট, স্বয়ং, হইতে, হইবে, হইয়া, হওয়া, হওয়ায়, হওয়ার, হচ্ছে, হত, হতে, হতেই, হন, হবে, হবেন, হয়, হয়তো, হয়নি, হয়ে, হয়েই, হয়েছিল, হয়েছে, হাজার, হয়েছেন, হল, হলে, হলেই, হলেও, হলো, হিসাবে, হিসেবে, হৈলে, হোক, হয়, হয়ে, হয়েছে, হৈতে, হইয়া, হয়েছিল, হয়নি, হয়েই, হয়তো, হওয়া, হওয়ার, হওয়ায়, ই, এক, এব, কমনে, কেথা, ভনজন, ধামার, পাচ, পি, পেয়, প্রযন্ত, বি, যাওয়া, র, সি, অর্থাত, আই, এমনকী, এল, তিনঐ, তাঁাহারা, তারৈ, নেওয়া, হয়েছেন, অন্যওজে, জে)

Figure 5.1: Stop Word

5.1.2 Filtered Search Key:

INPUT: বাংলাদেশ ক্রিকেট দল এর

After Removing Stop Word: বাংলাদেশ ক্রিকেট দল

Figure 5.2: INPUT

5.1.3 Retrieved Documents After Filtering:

['বাংলাদেশ ক্রিকেট দল', 'বাংলাদেশ ক্রিকেট দলের হেড', 'বাংলাদেশ ক্রিকেট দলের হেড কোচ', 'বাংলাদেশ ক্রিকেট দলের ওয়ানডে', 'বাংলাদেশ ক্রিকেট দলের', 'বাংলাদেশ ক্রিকেট দলের অনুশীলন', 'বাংলাদেশ ক্রিকেট দলের সহকারী', 'বাংলাদেশ ক্রিকেট দলের সহকারী কোচ', 'বাংলাদেশ ক্রিকেট দলের স্পন্সরশিপ', 'বাংলাদেশ ক্রিকেট দলের স্পন্সরশিপ স্বত্বাধিকারী', 'বাংলাদেশ ক্রিকেট দলের সাফল্যের', 'বাংলাদেশ ক্রিকেট দলের সাফল্যের গ্রাফটা', 'বাংলাদেশ ক্রিকেট দলের নৈপুণ্যটা', 'বাংলাদেশ ক্রিকেট দলের অস্ট্রেলিয়া', 'বাংলাদেশ ক্রিকেট দলের অস্ট্রেলিয়া সফর', 'বাংলাদেশ ক্রিকেট দলকে খালেদার', 'বাংলাদেশ ক্রিকেট দলকে খালেদার অভিনন্দন', 'বাংলাদেশ ক্রিকেট দল নিউজিল্যান্ডের', 'বাংলাদেশ ক্রিকেট দল নিউজিল্যান্ডের ওয়েংগিরি', 'বাংলাদেশ ক্রিকেট দলকে সংবর্ধনা', 'বাংলাদেশ ক্রিকেট দলের ম্যানেজার সাব্বির', 'বাংলাদেশ ক্রিকেট দলের ম্যানেজার', 'বাংলাদেশ ক্রিকেট দলকে', 'বাংলাদেশ ক্রিকেট দল নিউজিল্যান্ডে', 'বাংলাদেশ ক্রিকেট দলের ক্রিকেটারদের', 'বাংলাদেশ ক্রিকেট দলের ক্রিকেটারদের ম্যাচ', 'বাংলাদেশ ক্রিকেট দলের সাবেক', 'বাংলাদেশ ক্রিকেট দলের সাবেক অধিনায়ক', 'বাংলাদেশ ক্রিকেট দলকে অভিনন্দন বঙ্গবীর', 'বাংলাদেশ ক্রিকেট দল অস্ট্রেলিয়ায়', 'বাংলাদেশ ক্রিকেট দলের ম্যাচের', 'বাংলাদেশ ক্রিকেট দলের লক্ষ্যন', 'বাংলাদেশ ক্রিকেট দলের দায়িত্ব', 'বাংলাদেশ ক্রিকেট দলের হাই', 'বাংলাদেশ ক্রিকেট দলের অধিনায়কত্ব', 'বাংলাদেশ ক্রিকেট দলের অধিনায়কত্ব পেয়েছিলেন', 'বাংলাদেশ ক্রিকেট দলে সদ্য সাবেক', 'বাংলাদেশ ক্রিকেট দলে', 'বাংলাদেশ ক্রিকেট দলে সদ্য', 'বাংলাদেশ ক্রিকেট দলের কয়েকজন', 'বাংলাদেশ ক্রিকেট দলের প্রধান', 'বাংলাদেশ ক্রিকেট দলের প্রধান কোচ', 'বাংলাদেশ ক্রিকেট দলের কোচের', 'বাংলাদেশ ক্রিকেট দলের কোচের দায়িত্ব', 'বাংলাদেশ ক্রিকেট দলের অধিনায়ক', 'বাংলাদেশ ক্রিকেট দলের অধিনায়ক মাশরাফি', 'বাংলাদেশ ক্রিকেট দলের কোচ', 'বাংলাদেশ ক্রিকেট দলকে অভিনন্দন', 'বাংলাদেশ ক্রিকেট দলকে অভিনন্দন জানিয়েছেন', 'বাংলাদেশ ক্রিকেট দলের সকল']

Figure 5.3: Filtered Documents

5.1.4 Top 25 Documents After Applying BM25 Algorithm:

0 ->> বাংলাদেশ ক্রিকেট দলকে অভিনন্দন জানিয়েছেন ->> -0.29161532900934084
1 ->> বাংলাদেশ ক্রিকেট দল নিউজিল্যান্ডের ওয়েংগিরি ->> -0.29455705592826875
2 ->> বাংলাদেশ ক্রিকেট দলের স্পন্সরশিপ স্বত্বাধিকারী ->> -0.29697323269648135
3 ->> বাংলাদেশ ক্রিকেট দলকে অভিনন্দন বঙ্গবীর ->> -0.29721990856362585
4 ->> বাংলাদেশ ক্রিকেট দল নিউজিল্যান্ডে ->> -0.2974443997770639
5 ->> বাংলাদেশ ক্রিকেট দলের অধিনায়কত্ব পেয়েছিলেন ->> -0.2976949269620591
6 ->> বাংলাদেশ ক্রিকেট দলের নৈপুণ্যটা ->> -0.29779154890604503
7 ->> বাংলাদেশ ক্রিকেট দলের অনুশীলন ->> -0.29888618609247336
8 ->> বাংলাদেশ ক্রিকেট দলের সহকারী কোচ ->> -0.29897350375340814
9 ->> বাংলাদেশ ক্রিকেট দল নিউজিল্যান্ডের ->> -0.29931736206073023
10 ->> বাংলাদেশ ক্রিকেট দলের স্পন্সরশিপ ->> -0.29949687497609545
11 ->> বাংলাদেশ ক্রিকেট দলের হেড কোচ ->> -0.30001182360804435
12 ->> বাংলাদেশ ক্রিকেট দলের ওয়ানডে ->> -0.30006396768222315
13 ->> বাংলাদেশ ক্রিকেট দলের অধিনায়ক মাশরাফি ->> -0.30012927373949627
14 ->> বাংলাদেশ ক্রিকেট দলের প্রধান কোচ ->> -0.30048884598807796
15 ->> বাংলাদেশ ক্রিকেট দলকে অভিনন্দন ->> -0.3006346836528616
16 ->> বাংলাদেশ ক্রিকেট দলের ম্যানেজার ->> -0.30129168705567516
17 ->> বাংলাদেশ ক্রিকেট দলের ম্যানেজার সাব্বির ->> -0.30220028910437596
18 ->> বাংলাদেশ ক্রিকেট দলের সাফল্যের গ্রাফটা ->> -0.30238324356657165
19 ->> বাংলাদেশ ক্রিকেট দলের কয়েকজন ->> -0.3027412360079192
20 ->> বাংলাদেশ ক্রিকেট দলের অধিনায়ক ->> -0.30332801633496753
21 ->> বাংলাদেশ ক্রিকেট দলের ম্যাচের ->> -0.3038196087576372
22 ->> বাংলাদেশ ক্রিকেট দলের সাবেক অধিনায়ক ->> -0.3038521568543698
23 ->> বাংলাদেশ ক্রিকেট দলের প্রধান ->> -0.30387374160721786
24 ->> বাংলাদেশ ক্রিকেট দলের অস্ট্রেলিয়া সফর ->> -0.303927355509451

Figure 5.4: Ranked Documents

5.1.5 Collected Additional Terms From Top Documents:

(অভিনন্দন জানিয়েছেন নিউজিল্যান্ডের ওয়েংগিরি স্পন্সরশিপ স্বত্বাধিকারী বঙ্গবীর নিউজিল্যান্ডে অধিনায়কত্ব পেয়েছিলেন নৈপুণ্যটা অনুশীলন সহকারী কোচ হেড ওয়ানডে অধিনায়ক মাশরাফি প্রধান ম্যানেজার সাব্বির সাফল্যের গ্রাফটা কয়েকজন ম্যাচের সাবেক অস্ট্রেলিয়া সফর)

Figure 5.5: Additional Terms

5.1.6 Final Output:

(বাংলাদেশ OR পাকিস্তান OR এশিয়ান OR জাপান OR বিশ্ব) AND (ক্রিকেট OR ফুটবল OR হকি OR আইপিএল OR বিপিএল OR ফুটবলের OR টিটোয়েন্টি) AND (দল) OR (অভিনন্দন OR ধন্যবাদ OR শুভেচ্ছা OR স্বাগত OR সাধুবাদ OR মোবারকবাদ OR অভ্যর্থনা OR অভিবাদন OR শুভকামনা OR াগত) OR (জানিয়েছেন OR জানান OR জানালেন OR জানাচ্ছেন OR জানালেও OR জানায়) OR (নিউজিল্যান্ডের OR ইংল্যান্ডের OR অস্ট্রেলিয়ার OR জিম্বাবুয়ের OR শ্রীলঙ্কার OR আয়ারল্যান্ডের OR জিম্বাবুইয়ের OR অস্ট্রেলিযার OR নামিবিয়ার OR শ্রীলংকার OR কিউইদের) OR ( ) OR (স্পন্সরশিপ OR স্পনসর OR ফ্র্যাঞ্চাইজি OR ফ্রাঞ্চাইজি OR কোকাকোলা OR প্রিমিয়াম OR মেম্বারশিপ) OR (স্বত্বাধিকারী OR কর্ণধার OR মালিক OR প্রোপ্রাইটর OR প্রোপ্রাইটার) OR (বঙ্গবীর OR বীরউত্তম OR বীর OR সিদ্দিকী OR কৃষকশ্রমিক OR বিকল্পধারা OR বিকল্পধারার OR বীরোত্তম OR জেএসডি OR সেনানীদের OR মজলুম) OR (নিউজিল্যান্ডে OR অস্ট্রেলিয়ায় OR শ্রীলঙ্কায় OR অস্ট্রেলিয়াতে OR ইংল্যান্ডে OR জিম্বাবুয়েতে OR অস্ট্রেলিযায় OR ব্রিসবেনে OR ত্রিনিদাদে OR সিডনিতে OR মেলবোর্নে) OR (অধিনায়কত্ব OR ক্যাপ্টেন্সি OR আম্পায়ারিং OR উইকেটকিপিং OR স্লেজিং OR ফিল্ডিং) OR (পেয়েছিলেন OR পেলেন OR পেয়েছেন OR পান OR পেতেন OR পাচ্ছেন OR পাচ্ছিলেন) OR (নৈপুণ্যটা OR পারফরম্যান্সও OR পরিসংখ্যানটা OR পারফরম্যান্সটা OR পারফরমেন্সও OR ফর্মও OR পারফরমেন্সই OR ফর্মই) OR (অনুশীলন OR প্র্যাকটিস OR ওয়ার্মআপ OR কন্ডিশনিং OR পারফর্ম OR ফিল্ডিং OR প্রাকটিস) OR (সহকারী OR মোঃ OR ডেপুটি OR উপ OR অ্যাসিস্ট্যান্ট OR যুগ্ম OR এ্যাসিস্ট্যান্ট) OR (কোচ OR কোচ OR অধিনায়ক OR ট্রেনার OR অধিনায়ক OR হোয়াটমোর OR মিডফিল্ডার OR স্ট্রাইকার) OR (হেড OR ডিরেক্টর OR ম্যানেজার OR ডাইরেক্টর OR সেলস OR চিফ OR চীফ OR কমিউনিকেশন্স OR কোঅর্ডিনেটর OR কন্ট্রোলার OR ম্যানাজার) OR (ওয়ানডে OR টেস্ট OR টিটোয়েন্টি OR ওডিআই OR টি২০ OR টেস্টে OR টোয়েন্টি২০) OR (অধিনায়ক OR কোচ OR দলনায়ক OR দলপতি OR কোচ OR অলরাউন্ডার) OR (মাশরাফি OR মুশফিক OR সাকিব OR তামিম OR ধোনি OR মুস্তাফিজ OR মাহমুদউল্লাহ OR মুশফিকও OR মাহমুদুল্লাহ OR মিসবাহ) OR (প্রধান OR জ্যেষ্ঠ OR চিফ OR পরিচালক OR চীফ) OR (ম্যানেজার OR ডিরেক্টর OR ব্যবস্থাপক OR মহাব্যবস্থাপক OR ডাইরেক্টর OR কোঅর্ডিনেটর OR সিইও OR পরিচালক OR উপমহাব্যবস্থাপক) OR (সাব্বির OR রুম্মন OR জুবায়ের OR তাইবুর OR মাইশুকুর OR রুম্মান OR সাবিবর OR সোহান OR আশিক) OR (সাফল্যের OR সফলতার OR নৈপুণ্যের OR জয়ের OR কৃতিত্বের OR পারফরমেন্সের OR পারফরম্যান্সের OR অগ্রগতির OR পারফর্মেন্সের OR উন্নতির) OR (গ্রাফটা OR ধারাটা OR ধারাবাহিকতাও OR পারদটা OR গতিটা OR পাল্লাটা OR চিত্রটা OR মানটা OR স্বর্ণশিখরে OR ধারাবাহিকতাটা) OR (কয়েকজন OR কিছুসংখ্যক OR চারপাঁচজন OR তিনচারজন OR সাতআটজন OR পাঁচছয়জন OR কজন OR দুতিনজন OR দুএকজন) OR (ম্যাচের OR ম্যাচটির OR টেস্টের OR ম্যাচটার OR ওয়ানডের OR ফাইনালের OR সেমিফাইনালের OR ম্যাচটা) OR (সাবেক OR প্রাক্তন OR তৎকালীন OR জ্যেষ্ঠ OR মো OR তত্কালীন OR সিনিয়র OR মোহাম্মদ OR প্রয়াত OR প্রতিষ্ঠাতা) OR (অস্ট্রেলিয়া OR নিউজিল্যান্ড OR ইংল্যান্ড OR শ্রীলঙ্কা OR জিম্বাবুয়ে OR আয়ারল্যান্ড OR শ্রীলংকা OR হল্যান্ড OR আফ্রিকা OR কেনিয়া) OR (সফর OR বৈঠক)

Figure 5.6: Expanded Query

5.2 Analysis

From the result, we observe that:

1. All query expansion methods significantly outperformed the traditional BM25 method. Those explain that query expansion method is effective for overcoming lexical gap problem. They offers the improvement of performance for document ranking method.

2. Query expansion based on word embedding methods (QE-CBOW) exceed the external lexical methods. This explains that selecting additional word from semantic vector space is more effective than all other baseline process.

So, we can say that our methods have achieved the best performance compared with all the other baseline methods.

Chapter 6

Conclusion

In the following section we shall discuss about some of them and also determine our future works.

6.1 Discussion

In this paper, we proposed a new method for query expansion method for selecting additional word from a corpus and selecting similar words according to their semantic similarity with a word. The original queries and the expansion words are then used to rank documents to improve the performance of information retrieval tasks.

In future work, we hope to further improve our method by adding additional re-weighting method for similar words to re-rank the expansion words to improve the performance of information retrieval tasks.

References

[1] Deep, S. and Chawra, V., 2017, September. Improving search engine query expansion using clustering and indexing based approach. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 1836-1839). IEEE.

[2] Ganguly, D., Leveling, J. and Jones, G.J., 2010. Exploring sentence level query expansion in language modeling based information retrieval.

[3] Xiong, C. and Callan, J., 2015, September. Query expansion with Freebase. In Proceedings of the 2015 international conference on the theory of information retrieval (pp. 111-120). ACM.

[4] Singh, J., Prasad, M., Daraghmi, Y.A., Tiwari, P., Yadav, P., Bharill, N., Pratama, M. and Saxena, A., 2017, November. Fuzzy logic hybrid model with semantic filtering approach for pseudo relevance feedback-based query expansion. In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on (pp. 1-7). IEEE.

[5] Roy, D., Paul, D., Mitra, M. and Garain, U., 2016. Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608.

[6] Jan A Botha and Phil Blunsom. 2014. Compositional Morphology for Word Representations and Language Modelling.. In ICML. 18991907.

[7] Fei Cai, Maarten de Rijke, and others. 2016. [A survey of query auto completion in information retrieval.] Foundations and Trendső in Information Retrieval 10, 4 (2016), 273363.

[8] Wang, Y., Huang, H. and Feng, C., 2017, April. "Query expansion based on a feedback concept model for microblog retrieval". In Proceedings of the 26th International Conference on World Wide Web (pp. 559-568). International World Wide Web Conferences Steering Committee.

[9] Stephen E. Robertson, Hugo Zaragoza, Michael J. Taylor. "Simple BM25 extension to multiple weighted fields". CIKM., 2004, pp.42-49.

[10] Florian Beil, Martin Ester, Xiaowei Xu. "Frequent term-based text clustering". KDD., 2002, pp.436-442.

[11] George A. Miller. "WordNet: A Lexical Database for English". Communications of the ACM., Vol. 38, No.11, pp.39-41, 1995.

[12] Guangyou Zhou, Tingting He, Jun Zhao, Po Hu. "Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering". ACL. , 2015, pp.250-259.

[13] Meng Zhang, Yang Liu, Huan-Bo Luan, Maosong Sun, Tatsuya Izuha, Jie Hao. "Building Earth Movers Distance on Bilingual Word Embeddings for Machine Translation". AAAI.,2016, pp.2870-2876.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781, 2013.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality". NIPS., 2013, pp.3111-3119.

[16] Firth, John Rupert. A synopsis of linguistic theory 1930-1955. "Studies in Linguistic Analysis", pp.132.

[17] Yoshua Bengio, Rjean Ducharme, Pascal Vincent, Christian Janvin. "A Neural Probabilistic Language Model". JMLR., vol.3: pp.1137-1155, 2003.

[18] Ronan Collobert, Jason Weston, Lon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel P. Kuksa. "Natural Language Processing (Almost) from Scratch". JMLR., vol.12: pp.2493-2537, 2011.

[19] Diaz, F., Mitra, B. and Craswell, N., 2016. Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891.

[20] Kuzi, S., Shtok, A. and Kurland, O., 2016, October. Query expansion using word embeddings. In Proceedings of the 25th ACM international on conference on information and knowledge management (pp. 1929-1932). ACM.

[21] Hsi-Ching Lin, Li-Hui Wang, Shyi-Ming Chen. "Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques". Expert Syst. Appl., vol.31(2), pp.397-405, 2006.