Thesis for the Degree of B.Sc. Engineering

Thesis Title

Author Name
Student ID: 20111...

Department of Computer Science and Engineering
Bangabandhu Sheikh Mujibur Rahman Science and Technology
University, Gopalganj, Bangladesh

December, 2015

Thesis for the Degree of B.Sc. Engineering

Thesis Title

Author Name
Student ID: 20111...

Department of Computer Science and Engineering
Bangabandhu Sheikh Mujibur Rahman Science and Technology
University, Gopalganj, Bangladesh

December, 2015

Thesis Title

by

Author Name

Student ID:

Supervised by

Supervisor Name

Submitted to the Department of Computer Science and Engineering of Bangabandhu Sheikh Mujibur Rahman Science and Technology University in partial fulfillment of the requirements for the degree of B.Sc. Engineering

Thesis Evaluation Committee:

Teacher Name 1 .....

Teacher Name 2 .....

Teacher Name 3 .....

Thesis Approval

Student's Name:

Student's ID:

Thesis Title: :

We the undersigned, recommend that the thesis completed by the student listed above, in partial fulfillment of B.Sc. Engineering degree requirements, be accepted by the Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology for deposit.

Supervisor Approval*

.....
Name of Supervisor
Designation of Supervisor

Additional Approvals (if requires)*

.....
Name of Supervisor
Designation of Supervisor

Departmental Approval

.....
Name of Head of the Department
Chairman, Department of Computer Science and Engineering

Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh

Dedicated to my parents, Mr. A

And

Mrs. B

Acknowledgment

At first I like to thank almighty Allah who gives me ability to perform the Thesis. Then I like to give many thanks to my thesis supervisor Md.Nesarul Hoque, Lecturer, Department of Computer Science and Engineering, who encouraged, supervised and supplied necessary requirements and guideline in performing this work. His inspiration for doing research on Bangla Document Ranking by using Term Frequency and Cosine Similarity let me capable to complete the Thesis. I am grateful to him for his unvarying encouragement and simulating ideas. I also like to give thanks to all my teachers and also that of my friends who gave me advice for the completeness of the thesis. Finally I like to special thank to my parents and all my well wishers for their encouragement and support along my study life.

Shraboni Afroz Samapti

December, 2016

Abstract

Bengali is one of the ten most spoken languages in the world, with almost 200 million speakers. Growing online resources reveal a clear need for Bengali language applications and retrieval systems. The development of the internet, there are huge number of Bangla new articles are published every day on the web from different sources and this amount is growing rapidly day by day. Therefore, people have not enough time to read each document with a specified time limit. Bangla document ranking handles this difficulty with an efficient manner. Although, there are more works have been done in English and other European languages, but there is no contribution are present in Bangla language to rank a document. In our thesis, we rank Bangla document using term frequency and cosine similarity. We take document as vector and query as vector. We measure term frequency each word in each document. Then finding out score using cosine similarity. Then sorted the scores and ranked the documents. We demonstrate the effectiveness of the technique for Bangla document ranking.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction

In a real retrieval application (e.g.,Web search), the retrieval results using the initial query given by the user may not be satisfactory to the user; often, the user would need to revise the query to improve the retrieval/ranking accuracy. For some information seeking activities, the user may modify his query several times for one information need. In such an interactive retrieval scenario, the information available to us is more than just the current user query and the collection of documents. This paper summaries a set of experiments with term weighting for documents, using the measurement of term importance within an entire document collection. Then taking query it finds the relevant documents, if it exists then score the documents and sorting the scores it finds out the ranked documents. It is efficient while searching any document suppose in google users write the word or sentence that is called query,my proposed process will compute the importance of word and rank the documents sequentially which were in maximum times in documents.

1.2 Background and Present State of the Problem

Bangla document ranking is an important aspect but no such research work had done on bangla language. Various methods and techniques have implied to rank documents of different country of different languages. In reference the importance of a term within the entire document collection has found, then number of term matches between a query and a document , they used two functions 1)the raw frequency

and 2)the log2 of frequency. Then doing normalization for document length they combined the mesures and rank the documents.In reference, they used term frequency and inverse term frequency,they compute the document ranking performance using precision and recall.In reference, they used MMR method for ranking documents. I have mentioned earlier, no research work yet not performed on bangle. As there are several techniques for document ranking is present, these have their own limitations. One of the big problem of these systems is the accuracy is not better. If we need to implement these systems in real time purpose, the accuracy should increase otherwise it will be inefficient. So, I tried to increase the accuracy.

1.3 Motivation and Aims

Today the field of natural language processing(NLP) is increasing. As a nation we have the historical background of language movement, which reminds us that everyone has the right to communicate using their own language. The growth in electronically available documents makes research and applications in automatic document ranking more and more important. Huge number of available documents in digital media makes it difficult to obtain the necessary information related to the needs of a user. In order to solve this issue, document ranking system can be used. Many work have done in various languages in differnet countries but there is no previous works in bangla on the basis of document ranking. It encourages me to do the work on bangla document ranking . Now my aim is to identify the document on the basis of importance of words in bangla so that while doing web search user can see the ranked document corresponding their given query. By using the ranking process, a user can decide if a document is related to his/her needs without reading whole document. Document ranking systems can be categorized as extractive or abstractive according to the way the ranking is created. In extractive ranking approaches, the goals are identifying most important concepts in the input and giving relevant documents found in the document set as an output. In abstractive ranking approaches, first the system understands the texts and then it creates ranking with its own words. The abstractive

ranking is similar to the way a person creates ranking. Abstractive ranking remains as a difficult task in natural language processing.

1.4 Objectives and Specific aims

From the above discussion it is found that bangla document ranking is an important thing in our country while reading any bangla documents or newspapers. The objectives of my research are to identify the relevant documents within the given user query. And the specific aim is to rank relevant document with corresponding query. The query is matched word by word in relevant documents and ranked documents with cosine similarity.

1.5 What is NLP & Document Ranking?

Natural Language Processing, NLP is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages.

1.5.1 Document Ranking

Document ranking is the process of ranking documents with document and query.

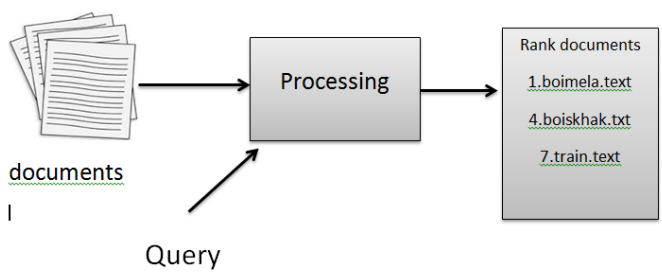

Figure 1.1: Main Process of our system

1.6 Organization of the Thesis

The dissertation is organized as follows:

- Chapter 1 Introduction. In this chapter an introduction to the periodic patterns mining researches is presented. The definition, importance and existing approaches are clearly introduced. After that, the dissertation focuses the contribution.

- Chapter 2 Related Work. This chapter first shows the state of the art methods of the periodic patterns mining research. Then describe two existing periodic pattern mining works $PSEMiner$ and $ListMiner$ in dynamic networks. The limitations of these methods are clearly addressed, as these are the focuses of this dissertation.

- Chapter 3 SPBMiner. We present our proposed technique for mining periodic behaviors in dynamic networks.

- Chapter 4 Experiments Analysis. In this chapter, it has been shown the effectiveness and efficiency of our proposed method.

- Chapter 5 Conclusion and Future Work. Finally, this chapter concludes the dissertation indicating the limitations and future works.

Chapter 2

Related Works

Chapter 3

Proposed System

We have used two methods for ranking document. One is term frequency and another is cosine similarity.

3.1 Creating Document Vector

The process of creating document vector is given below.

3.1.1 Term Frequency

A term that appears many times within a document is likely to be more important than a term that appears only once. D1: আমি বাংলাদেশকে ভালবাসি । বাংলাদেশ নদীমাতৃক দেশ।

D2: আমি বাংলাদেশের নাগরিক । কিন্তু বাংলাদেশের সকল নাগরিক তাদের মোলিক অধিকার পায় না।

D3: বাংলাদেশ উন্নয়নশীল দেশ। এখনো উন্নয়নের দিক থেকে পিছিয়ে আছে বাংলাদেশ ।

After performing the term frequency calculation we get the table:

3.1.2 Stop Word Removing

Documents contain words that do not add information but are necessary for syntactical formation,such as words like এবং , অথবা , কিন্তু etc. Since these words are less useful and less informative, they introduce noise into the document representation. In order to get rid of these kind of words, a stop word removal step is used. Stop word removal is done using predefined, human-made list of words. Since a predefined list is used, this approach is language dependent. Instead of using these kinds of lists, a

frequency threshold can be used. If a word is seen moreless frequently than predefined threshold, that word can be considered as stop word. But decision of threshold is another issue to be considered.

3.1.3 Stemming

In a document a word can be seen in different formays, such as plural vs. singular, present vs. past tense, etc. Most of the time these words have the same meaning and treating them differently is unnecessary. In order to use these words as the same token(concept), stemmers are used. Stremmers are tools that reduce the orginal word forms into roots(stems) of these words. Stemmers are necessary to represent different forms in a single format and to reduce memory usage for storing the words. Also, smaller list of words make it easir to perform calculations. As a result of performing stemming,document representation is less noisy and more dense. The efficiency of a stemmer is important while performing further calculations. Sometimes stemmers can do over-stemming such that two words are given the same stem,while it should not be. For example, the words "বাংলাদেশের" and "বাংলাদেশকে" are two different words, which should not be stemmed into the same root. But stemmers can find out their root as "বাংলাদেশ". Another steamming problem is related to under stemming such that two words should have been stemmed into the same word, but have not been. for example, "হাসি" and 'হাসানো' can be found as two different stems,instead of one.

3.1.4 Weighting Term

Then find weight in a document for each term using this equation

$$Weight++ = postings[term][doc] * IDF$$

3.1.4.1 Inverse Document Frequency(IDF)

A term that occurs in a few documents is likely to be a better discriminator than a term that appears in most or all documents.

Then assigning document IDS we keep them into database as vector after doing these steps.

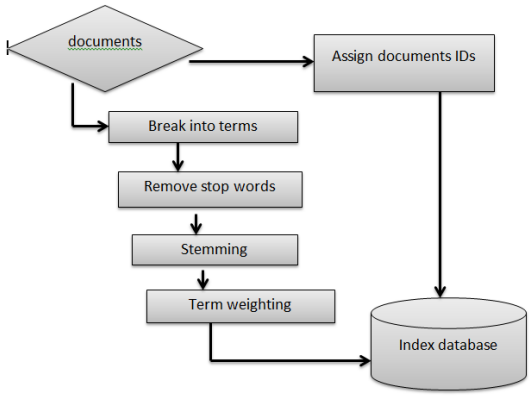

Figure 3.1: Indexing Document Ids & keep it in a vector.

3.2 Preprocessing of Taking Query

For taking search query the preprocessing steps have to be done. After taking query step by step process has been done . And then match query with relevent documents. The preprocessing steps are showing in the figure 3.2 below.

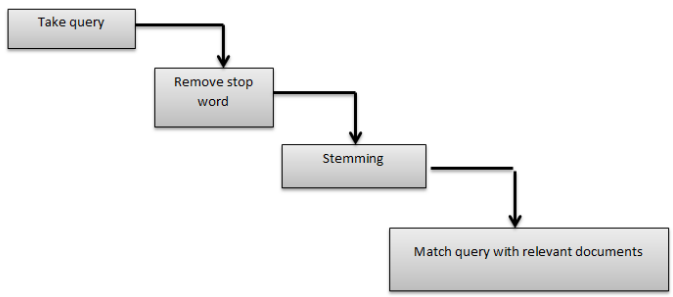

Figure 3.2: Pre processing of Input query

3.2.1 Removal of Stop Word

The less importance word should be removed from query. Sometimes input query contain words that do not add information, such type words ‘এবং’ , ‘অথবা’ , ‘কিন্তু ’ etc should remove from document.

3.2.2 Stemming

Finding root words from other similar words which have the same meaning and treating them differently is unnecessary.

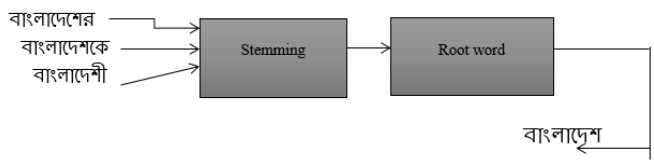

Figure 3.3: Stemming processing

3.3 Cosine Similarity

The query matches with relevent documents. If exists, find out the relevent documents and give them a scores using query term’s IDF multiplying with weight. Then sort the values and rank the documents.

Example: There are three documents.

D1: আমি বাংলাদেশকে ভালবাসি । বাংলাদেশ নদীমাতৃক দেশ।

D2: আমি বাংলাদেশের নাগরিক । কিন্তু বাংলাদেশের সকল নাগরিক তাদের মোলিক অধিকার পায় না।

D3: বাংলাদেশ উন্নয়নশীল দেশ। এখনো উন্নয়নের দিক থেকে পিছিয়ে আছে বাংলাদেশ ।

Input Query: বাংলাদেশের নাগরিক হিসেবে বাংলাদেশের উন্নতি চাই।

$$Cosine(document, query) = (document.query)/|documentvt||queryvt|$$

3.4 Main Approaches

When the preprocessing is completed, then our main process focuses on the analysis of the processed output by our own developed tools for Bangla document ranking. Our main approach are includes to generating term frequency vector, assigning cosine similarity to the corresponding document are known as document scoring. Then sorting the scores, we rank the documents.

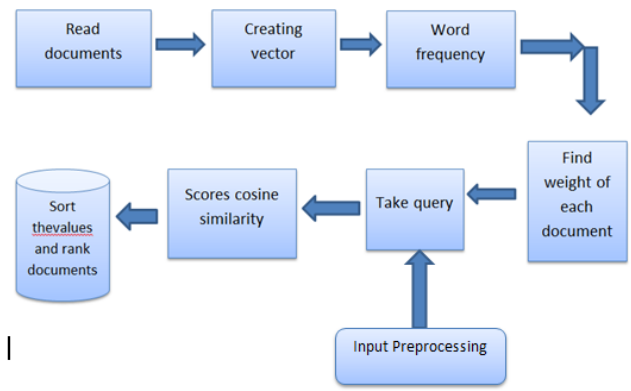

Figure 3.4: Overall graphical view of our proposed system

Chapter 4

Implementation

4.1 Bangla Corpus

Bangla language contains huge range of vocabulary which makes the language variegated in the world. Here, we develop a dictionary data set where 30 documents are used. In this system, we use the dictionary several times. It has two main reasons to access this dictionary. First one is to check a word which is rooted or not and second one is to get the associated POS tag for a word.

4.2 Bangla Stemming

Stemming is an operation that splits a word into its constituent root part and affix without doing complete morphological analysis. Terms with common stems tend to have similar meaning, which makes stemming an attractive option to increase the performance of news categorization task, where morphological analysis would be too computationally expensive. Another advantage of stemming is that it drastically reduce the vocabulary size of highly inflected languages corpus like Bangla. The algorithm of using bangla stemmer

Algorithm 1: bangla-stemmer (word)

```
while each word 2 document do
    dictionary-checkers(word);
    if word 2 dictionary then
        stem=word;
    else
        stem1=stemming(word);
        if stem1 2 dictionary then
            stem=stem1
        else
            stem=stemming(stem1)
        end
    end
end
```

4.3 Bangla Stop Words

Statistical analysis through the documents showed that some words have quite low frequency, while some others act just the opposite. The common characteristic of these words is that they carry no significant information and used just because of grammar. This set of words are usually known as stop words. In the resulting stop word list, there were thus a large number of pronouns, articles, prepositions, and conjunctions. As in various English stop-word lists, there were also some verbal forms. When using, this stop word list, the vocabulary size reduced significantly.

4.4 Term Weighting

Term Weighting of documents can be evaluated by measuring the Term Frequency (TF) and Inverse Document Frequency (IDF). These are the statistical measurement of weight that is intended to determine the importance of a word for a document in

a corpus. It can be used for stop-word filtering. This is the combine definition of Normalized Term Frequency and Inverse Document Frequency.

4.4.1 TF

Term Frequency measures how frequently a word occurs in a document. Different document varies in length. Therefore, a word can be occurring more times in a larger document than shorter. Thus, the raw frequency of a word is divided by the length of the document.

$tf(t,d) = \frac{f(t,d)}{lengthofd}$

here $f(t,d)$ is the raw frequency of word t in document d.

4.4.2 IDF

Inverse Document Frequency measures the importance of a word within a document. TF provides same importance for every word, where IDF provides less weight to the frequent word and high weight to the rare word.

$idf(t) = log(N/DF)$

The TF-IDF weighting scheme sets a weight to a word t in document d

$tf - idf(t,d) = tf(t,d) * idf(t)$

The weight of a word t in document d is highest when t occurs many times in a small number of document, and lower when t occurs a very few times in a document or occurs in many documents of the corpus.

4.5 Cosine Similarity between query and document

The similarity measures comparing between document vector and query vector. Though the angle between two vectors considered (0 to 90), than the similarity lies between 1 to 0.

$Similarity = \cos\theta = \frac{\vec{A}.\vec{B}}{|\vec{A}||\vec{B}|}$

Chapter 5

Experimental Result & Discussion

5.1 Experimental Result

To examine our ranking, we collect 30 bangla documents from the bangla daily news-paper eg Prothom alo, kaler kontho etc. The documents are written and saved in the text files using UTF-8 format. For each document in our corpus, we consider only one human ranking for evaluation. Evaluation of a system produced ranking is done by comparing it to the human ranking. There are some documents here on boimela.

Doc 1 : boimela txt.

কলকাতায় বাংলাদেশ বইমেলা থাকছে ৫০ প্রকাশনী । বইয়ের বন্ধুত্ব সীমানা ছাড়িয়ে- স্লোগানে ১ সেপ্টেম্বর থেকে কলকাতায় শুরু হচ্ছে ১০ দিনব্যাপী 'বাংলাদেশ বইমেলা'। সচিবালয়ে সোমবার সংস্কৃতি সচিব আকতারী মমতাজ এক সংবাদ সম্মেলনে জানান, ষষ্ঠবারের মতো আয়োজিত মেলায় বাংলাদেশের ৫০টি প্রকাশনা প্রতিষ্ঠান অংশ নেবে। ১ সেপ্টেম্বর বিকেল ৫টায় মেলার উদ্বোধন করবেন ইমেরিটাস অধ্যাপক আনিসুজ্জামান। উদ্বোধনী অনুষ্ঠানে থাকবেন পশ্চিমবঙ্গ সরকারের শিক্ষামন্ত্রী পার্থ চট্টোপাধ্যায় ও কবি-প্রাবন্ধিক শঙ্খ ঘোষ। প্রতিদিন দুপুর ২টা থেকে রাত ৮টা পর্যন্ত মেলা উন্মুক্ত থাকবে। শনি ও রোববার বিকাল ৩টা থেকে রাত ৮টা পর্যন্ত মেলা চলবে। জাতীয় গ্রন্থকেন্দ্র ও রপ্তানী উন্নয়ন ব্যুরোর সহযোগিতায় এবং কলকাতায় বাংলাদেশ উপ-দূতাবাসের ব্যবস্থাপনায় বাংলাদেশ জ্ঞান ও সৃজনশীল প্রকাশক সমিতি গত পাঁচ বছর ধরে কলকাতায় বাংলাদেশ বইমেলার আয়োজন করছে।প্রথম তিন বছর মেলাটি গণকেন্দ্র শিল্প সংগ্রহশালায় হলেও গত দুই বছর ধরে রবীন্দ্র সদনের উন্মুক্ত প্রাঙ্গণে হয়। এবারও এই উন্মুক্ত প্রাঙ্গণে বাংলাদেশ বইমেলা বলে জানান আকতারী মমতাজ।

Chapter 6

Conclusions and Future Work

Bibliography