BENGALI INFORMATION RETRIEVAL SYSTEM (BIRS)

Md. Kowsher[1], Imran Hossen[2] and SkShohorab Ahmed[2]

[1]Department of applied mathematics, Noakhali Science and Technology University, Noakhali-3814, Bangladesh
ga.kowsher@gmail.com
[2]Department of Information and Communication Engineering, University of Rajshai, Rajshai-6205, Bangladesh
imranhsobuj97@gmail.com
shohorab.ahmed.it@gmail.com

ABSTRACT

Information Retrieval System is an effective process that helps a user to trace relevant information by Natural Language Processing (NLP). In this research paper, we have presented present an algorithmic Information Retrieval System(BIRS) based on information and the system is significant mathematically and statistically. This paper is demonstrated by two algorithms for finding out the lemmatization of Bengali words such as Trie and Dictionary Based Search by Removing Affix (DBSRA) as well as compared with Edit Distance for the exact lemmatization. We have presented the Bengali Anaphora resolution system using the Hobbs' algorithm to get the correct expression of information. As the actions of questions answering algorithms, the TF-IDF and Cosine Similarity are developed to find out the accurate answer from the documents. In this study, we have introduced a Bengali Language Toolkit (BLTK) and Bengali Language Expression (BRE) that make the easiest implication of our task. We have also developed Bengali root word's corpus, synonym word's corpus, stop word's corpus and gathered 672 articles from the popular Bengali newspapers 'The Daily Prothom Alo' which is our inserted information. For testing this system, we have created 19335 questions from the introduced information and got 97.22% accurate answer.

KEYWORDS

Bangla language Processing, Information retrieval, Corpus, Mathematics, and Statistics.

1. INTRODUCTION

Information Retrieval (IR) simply refers to retrieve information from a collection of sources based on relevant query. It is a science for searching information in a document, searching documents themselves and searching of text, images and sounds . Searching is mainly either based on full-text or content-based. By default, IR means text retrieval because of historical reasons [1]. Recommender System that is intimately related to the Information Retrieval System work without a query.

Automated Information Retrieval are most likely used to reduce information overload. It is basically a software environment that gives access to documents like books, journals, magazines and so on. Documents can be stored and managed using this system. And today, web search engines are the most visible IR application.

Everyday, a huge number of information are produced by newspapers, social networking sites and different kinds of websites. Due to these large collections of digital documents in the web or local machine, finding the desired information is a tedious process. Finding relevant information based on query, has some challenges such as word mismatch that is a sentence can be made in different ways, their meaning is same but structure is different and a question can be formulated in different ways utilizing synonymuos words. So it is very challenging and difficult task to retrieve the desired information.

The Boolean Retrieval Model was used in the first search engines and is still in using today [2]. Documents will be retrieved if they correspond exactly to the query terms but this does not generate the ranking since this model assumes all the documents in the retrieved set to be relevant. However, it is due to the lack of the appropriate ranking algorithm that is the major drawback of this approach. Therefore, the Vector Space Model [3] was suggested. The word weighting and similarity can be defined in this space between papers. Terms are usually weighted by frequency in the document (TF or Term Frequency) or normalized with respect to the number of documents where they appear (IDF, or Inverse Document Frequency).

Unlike the previously mentioned retrieval models, probabilistic models provide some levels of theoretical assurances that the models will perform well as long as the model assumptions are consistent with the data observed. The Principle of Probability Ranking or PRP[4] states that ranking documents by the probability of relevance will maximize precision at any given rank — assuming a document's relevance to a query is independent of other documents. One of the first techniques was suggested in[5] to calculate the probability of relevance using document terms. Other methods assess text statistics in papers and queries and then use them to construct the term weighting function [6]. BM25 ranking algorithm works well in different tasks [7]. More advanced methods such as the Relevance-Based Language Models (or Relevance Models for short, RM) are the best-performing text retrieval ranking techniques [8].

So we are mainly interested to deal the problems using TF- IDF and cosine similarity. TF was used to calculate the importance of every word in a sentence. IDF calculated the actual importance of the words in a document. Then we used cosine similarity to figure out the relationship between questions and sentences. Our main objective is to retrieve relevant information within a short time with great accuracy.

2. RELATED WORK

In foreign language for instance English language, there are many distinct fields such as Web information retrieval [9], the retrieval of the picture and video [10] or the content-based recommendation [11], text retrieval methods that have been commonly studied.

However, in Bangla language, the study of information retrieval isn't satisfactory. A joint correlation technique [12] is used to extract and recognize Bangla numbers from the paper picture. Bangla OCR by the Center for Research Bangla Language Processing (CRBLP) which converts written text or pictures into editable Unicode text [13], has weaknesses as it is for a restricted scope. However, greater efficiency has been accomplished by the OCR scheme for Bangla[14] and the web handwritten OCR for Bangla [15].

Much research is performed previously on summarization, such as extracting Bangla sentences for document summarization [16]. In addition, knowledge extraction from Bangla Blogs and News [17], Bangla text extraction from real images [18], Bangla sentiment analysis from micro-blogs, media channel and customer feedback portal [19] is widely studied.

Unlike these works, we present an information recovery system on Bangla language by the help of mathematics and statistics.

3. PROPOSED WORK

In this paper, we introduced a Bengali Information Retrieval Systems (BIRS) based on Bengali Natural Language Processing (BNLP). The procedure we adopted is isolated in three parts such as collecting informative documents , pre-processing data and finding relationships between informative documents and questions via the boost of TF-IDF model and cosine similarity. Firstly, we collectd five types of corpus such as Bengali root words, stop words, our informative documents, questions, synonym words . Secondly, we pre-processed our desired data. Finally, Cosine similarity was used to obtain the relationship between the questions and the answers. However, cosine similarity deals with vectors. In this case, the documents and questions were converted to vectors using the TF-IDF model. For better explanation, we take a bunch of information as paragraph along with two questions so that a workflow of information retrieval can be explained.

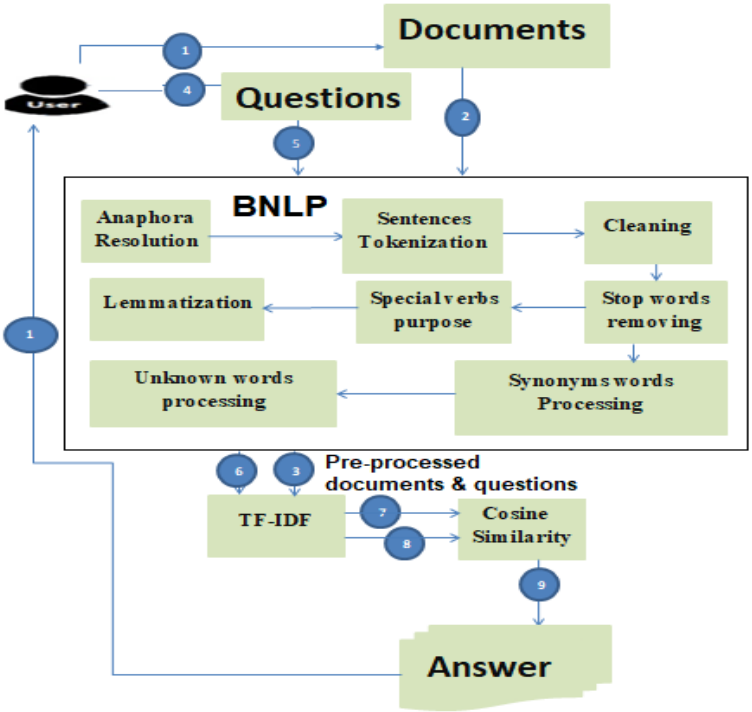

Fig.1: Proposed Work.

3.1. Category: Information

বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯২০ সালের ১৭ মার্চ টুঙ্গিপড়া গ্রামে জন্ম গ্রহণ করেন। তার রাজনৈতিক জীবন শুরু হয়েছিল ১৯৩৯ সালে মিশনারি স্কুলে পড়ার সময় থেকেই। ভারত পাকিস্থান বিভক্ত হয়ার পর পূর্ব পাকিস্থানের উপর পশ্চিম পাকিস্থানের অন্যায় অবিচার বাড়তে থাকে। এজন্য তিনি ১৯৬৬ সালের ৫ ফেব্রুয়ারি লাহোরে বিরোধী দলসমূহের একটি জাতীয় সম্মেলনে ঐতিহাসিক ছয় দফা দাবী পেশ করেন যা ছিল কার্যত পূর্ব পাকিস্তানের স্বায়ত্তশাসনের

পরিপূর্ণ রূপরেখা। অবশেষে তিনি ১৯৭১ সালের ২৬ মার্চ বাংলাদেশের স্বাধীনতার ঘোষণা দেন।

Category: Questions

Question-1: বঙ্গবন্ধু শেখ মুজিবুর রহমান কোথায় জন্ম গ্রহণ করেন?
Question-2: কত তারিখে তিনি ছয় দফা দাবী পেশ করেছিল?

3.2. Corpus

For the implication of Bengali Informative Retrieval System (BIRS), we mainly described five types of corpus. In the first corpus, there were 28,324 Bengali root words. The purpose of this corpus was to lemmatize Bengali words. The second one that contained 382 Bengali stop words was used to remove unnecessary information from documents e.g. stop words from the informative documents and questions. We compiled 672 articles as informative documents from variety field of interest such as politics, entertainment, sports, education, science, and technology from the popular Bengali newspapers named 'The Daily ProthomAlo' that was our third corpus as informative documents. In this work, as our fourth corpus, we created 19334 questions from our informative documents. Furthermore, for the sake of synonymous words processing, we included 18454 more Bengali similar words and collectively that was our fifth copus. However, there were also some other corpus e.g. verb processing, unknown word processing and removing punctuations and so on.

3.3. Pre-Processing

We need to pre-process or normalize the informative documents and the questions through cleaning, verb processing, removing stop words, tokenization, lemmatization, and synonyms words processing . Basically, the unwanted or special characters and stop-words don't affect on any linguistic operations. Besides, the stopwords of Bengali language always remains as root .So pre-processing these are also a good way for reducing execution time.

3.3.1. Anaphora

In linguistics, anaphora is the use of an expression whose interpretation depends upon another expression in context. In narrower sense, it refers to replace the previously described words (noun) with other words (pronoun) for further use in context. It requires a successful identification and resolution of Natural language processing(NLP). And hence presents challenges in computational linguistics.

In the proposed Bengali Informative Retrieval system (BIRS), we mentioned a review of work done in the field of anaphora resolution which has an influence on pronouns, mainly personal pronoun. The Hobbs' algorithm of Anaphora resolution was used in our proposed work. The algorithm has been adapted for Bengali language taking into account the roles of subject, object and its impact on anaphora resolution for reflexive and possessive pronouns. Here is the two sentences from our Bengali informative documents:

বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯২০ সালের ১৭ মার্চ টুঙ্গিপড়া গ্রামে জন্ম গ্রহণ করেন। তার রাজনৈতিক জীবন শুরু হয়েছিল ১৯৩৯ সালে মিশনারি স্কুলে পড়ার সময় থেকেই।
Here, 'তাঁর' (his) is the pronoun of 'বঙ্গবন্ধু শেখ মুজিবুর রহমান' (noun).

3.3.2. Tokenization

Tokenization is an NLP action. It separates a stream of text into words, sentences, symbols or phrases or other elements which are meaningful. We used the sentence tokenization by using the BLTK tool and we mapped every tokenized sentence to its root sentence.

3.3.3. Cleaning

Cleaning word refers to remove an unwanted character which doesn't sentiment to an informative documents. For example, colon, semicolon, comma, question mark, exclamation sign, and other punctuations don't provide meaningful connation. We used the Bangla Regular Expression (BRE) tool to shift the unwanted characters from the informative documents and the questions. The Bengali punctuation corpus was used to deal with BRE and carry away punctuation as unwanted data.

3.3.4. Stop Words Removing

Stop words refer the words that does not affect to the overall meaning of the sentences in the documents. For instance, in Bengali language stop words such as এবং (and), কোথায় (where),অথবা (or), তে (to), সাথে (with) are meaningless making the overall meaning of the sentences. As our proposed system i.e. BIRS is an algorithmic approach based on data, the stop words need to be dismissed. In our documents , every sentence was checked whether there was stopwords or not. If there was anstopword, the word was deleted. For the simplification of this action, we used Bengali Language Toolkit (BLTK).

3.3.5. Lemmatization for Bangla Language

Lemmatization is a significant process to transfer a word to its root word. It is one kind of Natural Language Processing technique and effective for use in various purposes such as text mining, answering questions or Chatbot etc.In Bengali natural language processing, there are few verbs that cannot be lemmatized by any system because of the limitation of lemmatization algorithms. For example, গেলে (went) and গিয়ে (going) are generated from the root word যাওয়া (go). There is no relation of character between গেলে(went) and যাওয়া (go). So processing these words with algorithms are not good choice. That is why, these types of verbs are mapped to their root verbs for easily accessing.

In our proposed BIRS , we used two novel techniques for lemmatization one is based on data structure and the other is the hash table of computer programming. The novel techniques are "DBSRA" and "Trie". Dictionary Based Search by Removing Affix (DBSRA) which is really an easy concept and more suitable for the lemmatization of Bengali words with the lowest time and space complexity. In this method, at first it removes the ith character from any data (which we want to shift into root word) where i =0, 1, 2,.... (Length of word - 1) and delete last jth character where j = length of word n, n-1, n-2, ..., 1 .After that, this method will look up in our corpus whether the intended words are in the root words corpus or not. "Trie" is a tree-based data structure which is exoteric for storing data. Like DBSRA, Trie also requires lowest space and time complexity. A Trie depot the inserted information's and question's words with the similar prefix under the identical sequence of edges in the tree eliminating that essential for storing the identical prefix each time for each word.

In BIRS, the Trie is taken into account as lemmatization procedure which accommodates for retrieving all possible lemmas where every node contains a single character and every two nodes are connected with a single edge. In our system, we used 'Trie' as an additional process

for the best lemma as sometimes, the Trie approach does not work properly if the input word contains a prefix. So eliminating the prefix is mandatory for proper lemmatization.

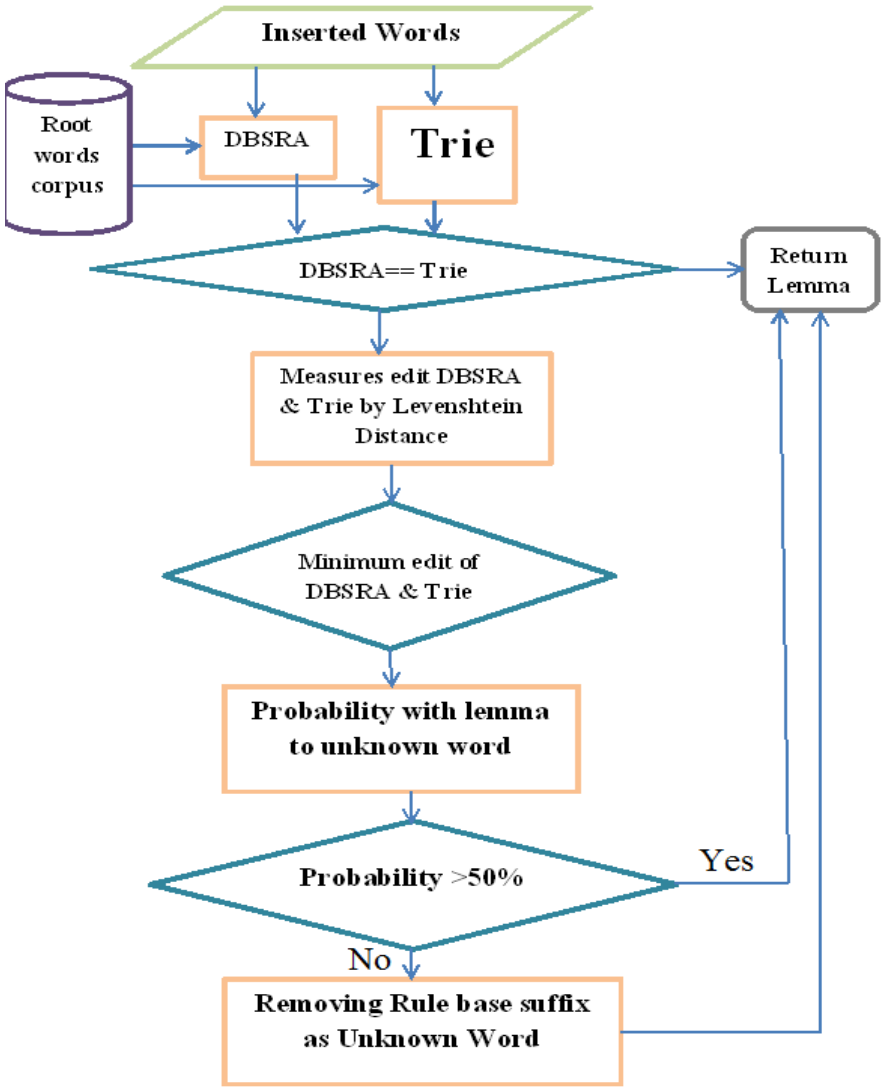

Fig.2. Bangla Lemmatization Process

For the best accuracy, the whole informative documents and questions words are processed with both lemmatization techniques. accurate lemma can be found out with the help of 'Edit distance'. Edit distance is egregious to count the similarity or dissimilarity between two words using dynamic programming.

In the lemmatization process, we applied "Edit Distance" for the root word comparison of 'DBSRA' and 'Trie'. In the applications of Natural Language Processing, Edit Distance is massively applied for the spelling correction of a word.

Sometimes, the lemmatization algorithms are not a good choice for the unknown words. Here, unknown words refer to the name of a place, a person or name of anything. 'Edit Distance' assists to determine which word is known or unknown. Edit Distance can count the probability with the edit to its word (not lemma). If the probability P(edit|word) is greater than 50% (P(edit|word) > 50%) then it is counted as unknown words.

In order to process the unknown words, we built a corpus, the suffix of Bengali Language that included তে (te), ছে(che), য়ের(yer), etc. The longest common suffix was removed from the last character of an unknown word obtained the lemma or root of an unknown word.

3.3.6. Synonyms Words Processing

Synonyms words refers same meaning with different words. There is a possibility when users make questions which are not available in information data but the meaning is same. In thi.s case, BIRS may be failed to answer the questions correctly. And hence the overall performance will be degraded if this type of event happens for any user. So we had given much importance in case of synonyms words processing in the BIQAS and also Natural Language Understanding. Therefore, to handle this unwanted situation, a synonyms word corpus were constructed containing total 13,189 words. Every word was mapped to a common identical word in the informative documents and th questions. In our desired system, if a word was not synonymous word with respect to the synonym corpus word, we didn't count it as a similar word.

After preprocessing the sentences and questions look the following:

Sentence-1: <s>বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯২০ সাল ১৭ মার্চ টুঙ্গিপড়া গ্রাম জন্ম গ্রহণ করা</s>
Sentence-2: <s>বঙ্গবন্ধু শেখ মুজিবুর রহমান রাজনীতি জীবন শুরু হয় ১৯৩৯ সাল মিশনারি স্কুল পড়া সময় থাকা</s>
Sentence-3:<s> ভারত পাকিস্থান বিভক্ত হয় পূর্ব পাকিস্থান পশ্চিম পাকিস্থান অন্যায় অবিচার বাড়া থাকা</s>
Sentecnce-4: <s>বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯৬৬ সাল ৫ ফেব্রুয়ারি লাহোর বিরোধ দল জাতী সম্মেলন ঐতিহাসিক ছয় দফা দাবী পেশ করা থাকা কার্য পূর্ব পাকিস্তান স্বায়ত্তশাসন পূর্ণ রূপরেখা</s>
Sentence-5: <s>অবশেষে বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯৭১ সাল ২৬ মার্চ বাংলাদেশ স্বাধীন ঘোষণা দেওয়া</s>

Question-1: বঙ্গবন্ধু শেখ মুজিবুর রহমান জন্ম গ্রহণ করা
Question-2: তারিখ বঙ্গবন্ধু শেখ মুজিবুর রহমান ছয় দফা দাবী পেশ করা

4. ALGORITHMS

4.1. TF-IDF

TF-IDF is an abbreviation for the Term Frequency-Inverse Document Frequency. Finding the importance of a word to a sentence is a numerical technique and is mathematically-statistically important.

In our proposed system of the TF-IDF model, there are several steps to figure out the TF-IDF of an inserted sentence. Actually, TF calculates the frequency of a term in a document. It refers how many times a term occurs in a document.

Firstly, the term rule has been ensured to measure the value of TF in every pre-processed sentence. For the standard data, normalization has been introduced due to the variation of the length of the document. In small document, there might be less word terms than the large

documents. Therefore there would be wrong in the retrieved information based on the query. And hence we introduced the normalization technique as a term (word÷length) of the sentence. Secondly, to figure out a relevant sentence by searching questions, IDF is pretty useful in this case. In TF all the words are treated as equal importance. But IDF determines the actual importance of a word in the document.

Finally, we counted the desired TF-IDF by multiplication with the term TF and IDF from the inserted questions and sentences. In order to reduce the time and the space complexity, we calculated the TF-IDF only of those words which are related to input questions inquired by the users.

Here, the TF-IDF from the previous example is shown in Table 1.

Table 1.TF-IDF of Sentences and Questions

Terms	Sent-1	Sent-2	Sent-3	Sent-4	Sent-5	Ques-1	Ques-2
টুঙ্গিপড়া	0.053767	0	0	0	0	0	0
রূপরেখা	0	0	0	0.026883	0	0	0
করা	0.007455	0	0	0.015305	0	0.056849	0.036176
রাজনীতি	0	0.046598	0	0	0	0	0
স্বায়ত্তশাসন	0	0	0	0.026883	0	0	0
স্বাধীন	0	0	0	0	0.053767	0	0
১৯৩৯	0	0.046598	0	0	0	0	0
ঐতিহাসিক	0	0	0	0.026883	0	0	0
সাল	0.030611	0.006461	0	0.003727	0.007455	0	0
৫	0	0	0	0.026883	0	0	0
থাকা	0	0.01479	0.018487	0.008533	0	0	0
অবশেষে	0	0	0	0	0.053767	0	0
পূর্ব	0	0	0.033162	0.015305	0	0	0
মিশনারি	0	0.046598	0	0	0	0	0
পশ্চিম	0	0	0.058248	0	0	0	0
ঘোষণা	0	0	0	0	0.053767	0	0
১৯২০	0.007455	0	0	0	0	0	0
বাংলাদেশ	0	0	0	0	0.053767	0	0
বঙ্গবন্ধু	0	0.006461	0	0.003727	0.007455	0.013844	0.00881
ফেব্রুয়ারি	0	0	0	0.026883	0	0	0
মার্চ	0.053767	0	0	0	0.030611	0	0
হয়	0	0	0.058248	0	0	0	0
শেখ	0.030611	0.006461	0	0.003727	0.007455	0.013844	0.00881
সম্মেলন	0	0	0	0.026883	0	0	0
পাকিস্থান	0	0	0.174743	0	0	0	0
শুরু	0	0.046598	0	0	0	0	0
দফা	0	0	0	0.026883	0	0	0.063543
১৭	0.007455	0	0	0	0	0	0
বাড়	0	0	0.058248	0	0	0	0
দাবী	0	0	0	0.026883	0	0	0.063543
পূর্ণ	0	0	0	0.026883	0	0	0
লাহোর	0	0	0	0.026883	0	0	0
জন্ম	0.053767	0	0	0	0	0.099853	0

পড়া	0	0.046598	0	0	0	0	0
ছয়	0	0	0	0.026883	0	0	0.063543
মুজিবুর	0.007455	0.006461	0	0.003727	0.007455	0.013844	0.00881
বিভক্ত	0	0	0.058248	0	0	0	0
অন্যায়	0	0	0.058248	0	0	0	0
পাকিস্তান	0	0	0	0.026883	0	0	0
জাতী	0	0	0	0.026883	0	0	0
সময়	0	0.046598	0	0	0	0	0
গ্রাম	0.007455	0	0	0	0	0	0
দেওয়া	0	0	0	0	0.053767	0	0
দল	0	0	0	0.026883	0	0	0
অবিচার	0	0	0.058248	0	0	0	0
গ্রহণ	0.053767	0	0	0	0	0.099853	0
স্কুল	0	0.046598	0	0	0	0	0
বিরোধ	0	0	0	0.026883	0	0	0
১৯৬৬	0	0	0	0.026883	0	0	0
জীবন	0	0.046598	0	0	0	0	0
ভারত	0	0	0.058248	0	0	0	0
পেশ	0	0	0	0.026883	0	0	0.063543
হয়	0	0.046598	0	0	0	0	0
কার্য	0	0	0	0.026883	0	0	0
১৯৭১	0	0	0	0	0.053767	0	0
২৬	0	0	0	0	0.053767	0	0
রহমান	0.053767	0.006461	0	0.003727	0.007455	0.013844	0.00881

4.2. Cosine Similarity

Cosine similarity is applied to obtain the relationship between questions. The cosine similarity is a measurement between two vectors that counts the cosine angle between them. The angle is the judgment of orientation but not magnitude that can be identified by comparing between vectors on a normalized space.

The cosine of two non-zero vectors can be derived as:

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

$$or, \cos\theta = \frac{\vec{A}.\vec{B}}{||A||||B||} = \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n} {A_i}^2} * \sqrt{\sum_{i=1}^{n} {B_i}^2}}$$

$$so, similarity = Cosine(question, document)$$

$$= \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n} {A_i}^2} * \sqrt{\sum_{i=1}^{n} {B_i}^2}}$$

With the help of the TF-IDF table, we calculated the cosine similarly which is shown in the table 2.

Table 2. Cosine Similarly

Relations	Sentence-1	Sentence-2	Sentence-3	Sentence-4	Sentence-5
Question-1	0.597	0.016	0.0	0.060	0.018
Question-2	0.074	0.012	0.0	0.483	0.013

The Cosine Similarity (Q1, S1) is greater than the Cosine Similarity (Q1, S2) (0.597>0.016). So it states, the percentage of question1's answer is 59.7% related to sentence-1 and 1.6% related to sentence-2. In the same way, the cosine similarity can be ordered from the table such as (Q1, S1) > (Q1, S4) > (Q1, S5) > (Q1, S2) > (Q1, S3). Similarly, the Cosine Similarity (Q2, S1) is less than the Cosine Similarity (Q2, S4) i.e. (0.074 < 0.483). So, the percentage of question2's answer is 7.4% related to sentence1 and 48.3% related to sentence-4. So the cosine similarity can be ordered from the table like (Q2, S4) > (Q2, S1) > (Q2, S5) > (Q2, S2) >(Q2, S3). Threfore, the answer of question-1 stays in sentence-1 and the answer of question-2 remains in sentence-4. In this way, we can find out the relevant answer from the corpus.

5. Experiments

5.1. Experimental Tools

The whole work was performed in Anaconda distribution and Python 3.6 tools. For the sake of clearing data, removing stop words, we have constructed a tool for Bengali language processing mentioned as Bengali language Toolkit (BLTK) and we have applied NLTK system in many pre-processing tasks.

5.2. Final Result

To test our proposed Bengali Information Retrieval System, we collected 672 articles from the popular Bengali newspapers ‘The Daily ProthomAlo’ as our input documents. We created 19334 questions for testing data and obtained 97.22 % accuracy of our BIRS. From 19334 questions, the number of correct answers was 18797 and incorrect was 537. The performance to retrieve information is pretty good and flexible with the lowest time complexity.

6. Conclusion And Future Work

In this paper, we have presented a Bengali Information Retrieval System. To establish our proposed system we have used various types of algorithms and methods such as lemmatization, anaphora resolution procedure, TF-IDF and Cosine Similarity. The whole actions have been processed with Bengali Language as part of BNLP. We have tested our proposed BIRS, noted the accuracy, compared the correct and incorrect results.

In future, we have a plan to improve the system for educational purposes, industry, business and personal tasks. And we want to use deep learning algorithms such as neural network for the development of the system.

REFERENCES

[1] Singhal, A. (2001). "Modern information retrieval: A brief overview.", IEEE Data EngineeringBulletin 24(4), 35–43.

[2] Croft, W.B., Metzler, D. &Strohman, T. (2009). " Search engines-information retrieval in practice.",Pearson education. http://www.search-engines-book.com/.

[3] Salton, G., Wong, A., & Yang, C. S. (1975). "A vector space model for automatic indexing." Communications of the ACM 18(11), 613–620. http://dx.doi.org/10.1145/361219.361220.

[4] Robertson &S.E. (1997) "Readings in information retrieval",The probability ranking principle in IR (pp. 281–286). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl. acm.org/citation.cfm?id=275537.275701.

[5] Robertson, S. E., & Jones, K. S. (1988) "Relevance weighting of search terms" (pp. 143–160). London, UK: Taylor Graham Publishing.

[6] Amati, G., & Van Rijsbergen, C. J. (2002). "Probabilistic models of information retrieval based on measuring the divergence from randomness." ACM Transactions on Information Systems 20(4), 357–389.

[7] Robertson, S. (2010). "The probabilistic relevance framework: BM25 and Beyond." Foundations and Trends in Information Retrieval 3(4), 333–389.

[8] Lavrenko, V., & Croft, W. B. (2001) "Relevance-based language models." In W. B. Croft, D. J. Harper, D.H.Kraft, &J.Zobel (eds.) SIGIR2001:Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, New Orleans, Louisiana, USA(pp.120–127). ACM.https://doi.org/10.1145/383952. 383972.

[9]. Agichtein, E., Brill, E., &Dumais, S. (2006) "Improving web search ranking by incorporating user behavior information." , Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 2006 (pp. 19–26). New York, NY, USA: ACM. https://doi.org/10.1145/1148170.1148177.

[10] Sivic, J., &Zisserman, A. (2003) "Videogoogle: A text retrieval approach to object matching in videos." ,Proceedings of the ninth IEEE international conference on computer vision, ICCV 2003 (Vol. 2, pp. 1470–1477). Washington, DC, USA: IEEE Computer Society. http://dl.acm. org/citation.cfm?id=946247.946751.

[11] Xu, S., Bao, S., Fei, B., Su, Z., & Yu, Y. (2008). "Exploring folksonomy for personalized search." , Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 2008 (pp. 155–162). New York, NY, USA: ACM

[12] M. K. I.Molla, & K. M.Talukder, (2007) "Bangla number extraction and recognition from the document image" , International Conference. on Computer and Information Technology, ICCIT 2007, pp. 512-517.

[13] M. S. Islam, (2009) "Research on Bangla Language Processing in Bangladesh: Progress and Challenges", International Conference on Language & Development pp. 23-25.

[14] M.A. Hasnat, S.M. Habib, & M. Khan (2008) "A high-performance domain specific OCR for Bangla script", Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics pp. 174-178, Springer, Dordrecht

[15] G. Fink, S. Vajda, U. Bhattacharya, S. K. Parui& B. B. Chaudhuri, (2010). " Online Bangla word recognition using sub-stroke level features and hidden Markov models" International Conference. on Frontiers in Handwriting Recognition, ICFHR 2010, pp. 393-398.

[16] K .Sarkar, (2012) "Bengali text summarization by sentence extraction", arXiv preprint arXiv:1201.224.

[17] A. Das & S. Bandyopadhyay, (2010)."Phrase-level Polarity Identification for Bengali" International Journal of Computational Linguistics and Applications, IJCLA, 1(1-2), pp. 169-182.

[18] U. Bhattacharya, S. K. Parui, & S. Mondal, (2009) "Devanagari and Bangla Text Extraction from Natural Scene Images", International Conference on Document Analysis and Recognition, pp. 171-175.

[19] A. Hassan, M.R. Amin, N. Mohammed, & A.K.A. Azad, (2016). "Sentiment Analysis on Bangla and Romanized Bangla Text (BRBT) using Deep Recurrent models", arXiv preprint arXiv:1610.00369