2018 2nd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET)
27-28 October 2018, Chittagong, Bangladesh

Vector Space Model based Topic Retrieval from Bengali Documents

Topu Dash Roy
Department of Computer Science and Engineering
North East University Bangladesh
Sylhet, Bangladesh
topucse05@gmail.com

Shamima Khatun
Department of Computer Science and Engineering
North East University Bangladesh
Sylhet, Bangladesh
shamimaneub07@gmail.com

Rubina Begum
Department of Computer Science and Engineering
North East University Bangladesh
Sylhet, Bangladesh
rubinarubi10@gmail.com

Al Mehdi Saadat Chowdhury
Department of Computer Science and Engineering
North East University Bangladesh
Sylhet, Bangladesh
amschowdhury@neub.edu.bd

***Abstract*—This work attempts to find the topic of a Bengali text document based on a traditional similarity based retrieval model named Vector Space Model. This fascinating model has traditionally obtained much fame in the research community, but to the best of our knowledge, was never tried for Bengali topic retrieval. In this work, therefore, we have used four different settings of the vector space model which are TF-IDF weighting scheme with Euclidean distance, TF-IDF weighting scheme with Manhattan distance, TF-IDF weighting scheme with Cosine similarity and Improved document scoring scheme. The K-nearest neighbor algorithm is then used to retrieve the topic of a query document. For training and testing purpose, we have also created a large corpus of Bengali text documents. On this corpus, our result shows the best retrieval accuracy of 93.33%.**

Keywords—Vector Space Model, TF-IDF, Cosine Similarity, Manhattan Distance, K-nearest Neighbor, Improved Document Scoring Scheme

I. INTRODUCTION

Topic retrieval (TR), a subfield of information retrieval (IR) system, deals with the problem of identifying the topic a document discusses about. Research on information retrieval is dated back to 1950[1] and still remains as a very important research field due to the need of information in various real world tasks including ranking web pages by search engines, document categorization, similar topic or document extraction and many more. Among many different IR tasks, the problem of retrieving topic from a document gained huge attention of the researchers for its inherent complexity and diverse application area.

A topic can be defined as a collection of similar terms that occurs frequently among documents. For example, terms like “medicine”, “doctor”, “hospital” occurs frequently when the topic is “Health”. This intuitive description leads to the formal problem formulation of the topic retrieval system. A topic retrieval problem assumes a collection of *m* documents available as a corpus *C*, that is, $C = \{D_1, D_2, ..., D_m\}$. Each document consists of a collection of *t* terms, $D_x = \{T_1, T_2, ..., T_t\}$, for any document *x*. It is also assumed that, the true topic of each of the document in the corpus is also known in advance. One or more query documents is then given to the retrieval system, on which prediction should be made. The query document also consists of a collection of terms $Q = \{T_1, T_2, ..., T_t\}$. Generally, a topic retrieval system compares this query document with the corpus by assigning some real number score $s(Q, D_i)$ to each i^{th} document of the corpus, which represents the similarity of the query with a particular document of the corpus.

On top of the formulated problem, several researchers has tried to solve the topic retrieval problem using several different techniques. All these techniques can be broadly categorized into several models including similarity based models which measures the similarity between a query document and documents from the corpus as the relevance criteria, probabilistic relevance models that estimates the probability of a binary random variable of success and failure, language models that looks at the semantic structure of the sentences and so on. Several interesting techniques including Probabilistic Latent Semantic Analysis[2], Latent Dirichlet Allocation[3] etc. has been proposed in the process.

Although diverse algorithms has been tried and tested for English language, topic retrieval in languages like Arabic [4,5,6], Tamil [7], Punjabi [8], Indonesian [9] etc. mostly took the text classification based retrieval approach. Almost all research work on topic retrieval done for Bengali language [10,11,12,13,14] primarily employed classification based approach. A major limitation in the classification based approach is that, they do not consider the semantic information present in a document. Moreover, the correlation between words were completely ignored. Past research on English suggests us that, it is a good idea to use an approach that uses the semantic information and word correlation information present in a document.

In this work, we therefore used a similarity based approach known as vector space model for the retrieval of the topic a Bengali document represents. To the best of our knowledge, this model has not ever being tried for Bengali topic retrieval. A vector space model is a matrix of weight values over a vocabulary, such weight values represents the significance of a particular word group in a topic. For our work, we’ve used TF-IDF weighting scheme and improved document scoring scheme for weighting criteria. Furthermore, the vector space model of the corpus is compared against a query document either directly or with

978-1-5386-8524-2/18/$31.00 ©2018 IEEE

similarity measurements such as Euclidean distance, Manhattan distance and cosine similarity to retrieve the topic from a query document. A retrieval accuracy of 93.33% has been found during the training phase.

Rest of the article is organized as follows. Section II describes the proposed dataset. Section III explains the whole process of using a vector space model with several weighting schemes, distance measures and K-nearest neighbor algorithm for retrieving topic of a document. Later in section IV, the result analysis on our proposed dataset is described. Section V concludes the article, which is followed by references of the work.

II. Preparing Dataset

The very first challenge in this research work was to collect a corpus of Bengali documents. Most of the corpus available in the web are either inappropriate for the task at hand or the size and the formatting of the corpus is not suitable for our work. Therefore, we decided to collect data on our own and make a new corpus suitable not only for the topic retrieval task, but also for other information extraction tasks as well.

A. Dataset Collection:

For this research work, we have collected a total of 9000 Bengali documents, belonging to nine categories namely "Science and Technology", "Health and Medicine", "Sports", "Entertainment", "Crime", "Tourism", "Recipe", "Religion" and "Economics". In our work, these categories are defined as topics that we want to retrieve. Each category contains 1000 documents and each document contains an average of 200 words. Most of these documents were collected from popular Bangladeshi newspaper site as prothom alo, kaler kantha, ittefaq, bdnews24, amar desh etc., popular blogs such as bioscope, ntv online and from several other magazines as well. A very small excerpt from one particular sample of the collected data belonging to the "Science and Technology" category can be seen from Fig 1.

দৈনিক ইত্তেফাক
০৫ আগষ্ট, ২০১৭ ইং ১৮:২৫ মিনিট
ক্রেতা-বিক্রেতার খুশিতে শেষ হচ্ছে স্মার্টফোন ও ট্যাব মেলা

আধুনিক স্মার্টফোন ক্রয়ে অভাবনীয় মূল্যছাড় ও উপহার পেতে ঢাকায় চলমান স্মার্টফোন মেলার শেষমুহূর্তে ক্রেতারা সারিবদ্ধ হচ্ছেন তাদের পছন্দের প্রতিষ্ঠানের স্টলে।রাজধানীর বঙ্গবন্ধু আন্তর্জাতিক সম্মেলন কেন্দ্রে শনিবার সকাল থেকেই মেলার প্রবেশমুখে দেখা যায় দর্শনার্থীদের ভিড়। গত বৃস্পতিবার থেকে শুরু হওয়া প্রযুক্তি পণ্য নিয়ে দেশের সবচেয়ে বড় আয়োজন টেকশহর স্মার্টফোন ও ট্যাব এক্সপোর পর্দা নামছে শনিবার রাত ৮টায়।

Fig. 1. Example of collected data from "Science and Technology" category

B. Dataset Preprocessing

In the data collection process, the dataset was kept noiseless. Thus broken lines, unstructured words, unnecessary white spaces and other unwanted texts were already absent from the dataset. For this work, we further preprocess our dataset to remove words, punctuations etc. which were irrelevant to the task of topic retrieval. Major steps in this preprocessing phase is explained below.

1) Stop-word Removal:

Stop words are words which are very common to any language that in general has almost no weights in retrieving topic of that document. Such words in English literature consists of in, at, for, to, am etc. We've created a list of 96 stop words for Bengali language. Some of which are given in Fig. 2.

এবং, হঠাৎ, তথাপি, নাহয়, সাথে, ওহে, আঃ, সুতরাং, ছিছি, পরক্ত, দিয়ে, গরগর, হয়তো, পুনশ্চ, না, অবশ্য, আচ্ছা, যথা, ওরে, কিন্তু, তাই, কতনা, অথবা, মোটকথা, বাহবা, সহসা, কী

Fig. 2. Some stop-words of Bengali Language

2) Digit and Punctuation Marks Removal:

Unlike other information retrieval tasks, in topic retrieval, digits and punctuation marks generally has very little significance. Therefore, those unwanted things were removed from the document.

3) Tokenization and Vocabulary Creation:

Similarity based topic retrieval models works with an assumption that, the semantic correlation between words in a document is overshadowed by the semantic correlation of words in a document to query. In other words, each word in a document can be thought of as a separate entity and comparing each term of the query with all terms of the document would represent the sematic correlation well. Therefore, in this work, we have tokenized all words of a document and from which we have created a vocabulary of all unique words across the whole corpus. This vocabulary acts as the basis for our vector space model.

III. Topic Retrieval From Bengali Documents using Vector Space Model

A vector space model (VSM) can simply be considered as a matrix of m rows and n columns where m represents the number of documents available in the corpus and n represents the size of the vocabulary created in the preprocessing phase. Such a matrix looks as Fig. 3. Values in each row corresponds to weights w_i assigned to each vocabulary term found inside the document. These weights forms the basis of further retrieval in this similarity based model.

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_m \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \cdots & V_n \end{array} \\ \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ w_{m,1} & \cdots & \cdots & w_{m,n} \end{bmatrix} \end{array}$$

Fig. 3. Vector Space Model

A. Term Weighting

There are several term weighting techniques available in existing research. In this work we had employed two of them as explained in the subsection below.

1) TF-IDF:

The full form of TF-IDF weighting is Term Frequency – Inverse Document Frequency. Term Frequency is a real number that represents the relative weight of a term in a particular document, whereas inverse document frequency penalizes this term frequency by reducing weight value for common terms that occurs throughout the corpus. The equation we used for obtaining the term frequency was:

$$TF(D_i,t) = \frac{\#\ of\ times\ term\ t\ appears\ in\ document\ D_i}{Total\ \#\ of\ terms\ in\ D_i} \quad (1)$$

The equation we used for obtaining the inverse document frequency was:

$$IDF(C,t) = \log_e\left(\frac{Total\ \#\ of\ documents\ in\ the\ corpus\ C}{\#\ of\ documents\ with\ term\ t\ in\ it}\right) \quad (2)$$

Both this TF and IDF quantity is then multiplied together to give final weight values of the vector space model. A special finding on TF-IDF weighting scheme is that, it does not consider query document's weights, therefore, a similarity measure will be required to compare the test query with the training corpus.

2) Improved Document Scoring Scheme:

This weighting technique taken from [15] is essentially a modified version of the TF-IDF weighting technique with the difference being the incorporation of query documents term frequency in the equation. The equation for this scheme is given below.

$$ImprovedScore = \sum_{t\in Q,D} \frac{(1+\ln(1+\ln(TF)))}{(1-s)+s\frac{DL}{AvDL}} \times QTF \times \ln\frac{m+1}{DF} \quad (3)$$

In this equation, *TF* is the regular term frequency in the document, *DL* is the length of the document with which the query is being compared, *AvDL* is the average document length throughout the corpus, *QTF* is the query term frequency, *m* is the number of documents present in the corpus, *DF* is the number of document that contains the term *t* and finally *s* is an empirical parameter that is needed to be found during experimentation. In our experiment, the most optimal value of *s* has been found to be *0.01*.

This equation is iterated by comparing one particular query Q, with each document D in the corpus one by one, taking each term common in both Q and D into consideration and finally all calculated weights are added to give the final score of a document based on the query. In the retrieval phase this score is used to rank documents, and from top k documents, the maximum occurred topic is returned as the target topic of the query.

B. Similarity Measures

Improved document scoring scheme directly compares the query document with the corpus and creates a vector space model based on this combined weights, therefore no similarity measure is needed for this weighting scheme. On the other hand, TF-IDF do no such things. Thus, similarity measurements are used to compare the vector space model created by TF-IDF scheme with the query document to return the target topic of the query document. Several similarity measurements are available in the literature. In this work we've used three of them as described below.

1) Euclidean Distance:

Euclidean distance is the most common distance measurement used in the literature over time. The basic equation for Euclidean distance measurement is:

$$E(Q,D_i) = \sqrt{\sum_{t\in Q,D_i}(Q_t - D_t)^2} \quad (4)$$

In this equation, Q_t is the weight of a particular term in the query which can be found by forming a TF-IDF weighting vector of the query document, and D_t is the weight of a particular document D_i, which is found from the vector space model.

2) Manhattan Distance:

Manhattan distance, another popular distance measurement has the equation:

$$M(Q,D_i) = \sum_{t\in Q,D_i} |Q_t - D_t| \quad (5)$$

Q_t and D_t is defined similarly as defined in the Euclidean distance measurement.

3) Cosine Similarity:

Cosine similarity measures the cosine angle between the query document and each training documents in the corpus. The value of this similarity measure ranges from 0 to 1, which makes it ideal for many weighting applications. The equation for cosine similarity measurement is:

$$\cos(Q,D_i) = \frac{Q.D_i}{|Q||D_i|} = \frac{\Sigma_{t\in Q,D_i} Q_t D_t}{\sqrt{\Sigma_{t\in Q} Q_t^2}\sqrt{\Sigma_{t\in D_i} D_t^2}} \quad (6)$$

Unlike Euclidean and Manhattan distance measure where fewer value means more similarity, in cosine measurement the higher the value means more similar a query to the document.

C. K-nearest Neighbor:

Both improved document scoring scheme and similarity measures returns a real number against each document representing the rank of a document based on the query. Therefore, we applied K-nearest Neighbor algorithm with K=35 to retrieve top 35 ranked documents based on the query and from these 35 documents, the maximum occurred topic is returned as the predicted topic of the query document.

IV. Results and Discussion

A total of four different approaches has been employed in this work as described earlier. Also a corpus of 9000 documents were collected for this work. From this corpus, 2790 documents were taken for training and testing purpose. These documents were distributed randomly making 2700 training documents and 90 test documents. The result of applying all four approaches in testing data is summarized in the confusion matrix given in Fig. 4.

	Sci & Tech	Health & Med	Sports	Entertainment	Crime	Tourism	Recipe	Religion	Economics
Sci & Tech	6	0	0	0	0	0	1	1	
Health & Med	0	7	0	0	0	1	2	0	0
Sports	0	0	6	0	0	0	0	0	
Entertainment	0	0	0	8	0	0	0	1	1
Crime	0	0	0	0	13	0	0	0	2
Tourism	0	0	0	0	0	14	0	0	0
Recipe	0	0	0	0	0	0	9	0	0
Religion	0	0	0	0	0	0	0	6	0
Economics	0	0	0	0	1	0	0	0	11

(a)

	Sci & Tech	Health & Med	Sports	Entertainment	Crime	Tourism	Recipe	Religion	Economics
Sci & Tech	0	0	0	0	0	8	0	0	0
Health & Med	0	0	0	0	0	10	0	0	0
Sports	0	0	0	0	0	5	0	1	0
Entertainment	0	0	0	0	0	9	0	1	0
Crime	0	0	0	0	0	15	0	0	0
Tourism	0	0	0	0	0	14	0	0	0
Recipe	0	0	0	0	0	2	7	0	0
Religion	0	0	0	0	0	0	0	6	0
Economics	0	0	0	0	0	6	0	5	1

(b)

	Sci & Tech	Health & Med	Sports	Entertainment	Crime	Tourism	Recipe	Religion	Economics
Sci & Tech	0	0	0	0	0	0	8	0	0
Health & Med	0	6	0	0	0	0	4	0	0
Sports	0	0	5	0	0	0	1	0	0
Entertainment	0	0	0	0	0	0	10	0	0
Crime	0	0	0	0	11	0	4	0	0
Tourism	0	0	0	0	0	8	6	0	0
Recipe	0	0	0	0	0	0	9	0	0
Religion	0	0	0	0	0	0	5	1	0
Economics	0	0	0	0	0	0	2	0	10

(c)

	Sci & Tech	Health & Med	Sports	Entertainment	Crime	Tourism	Recipe	Religion	Economics
Sci & Tech	8	0	0	0	0	0	0	0	0
Health & Med	0	8	0	0	0	0	2	0	0
Sports	0	0	6	0	0	0	0	0	0
Entertainment	0	0	0	10	0	0	0	0	0
Crime	0	0	0	0	13	1	0	0	1
Tourism	0	0	0	0	0	14	0	0	0
Recipe	0	0	0	0	0	0	9	0	0
Religion	0	0	0	0	0	0	0	6	0
Economics	0	0	0	0	2	0	0	0	10

(d)

Fig. 4. Confusion matrix for a) Improved Document Scoring Scheme b) Euclidean distance c) Manhattan Distance d) Cosine Similarity

Finally the accuracy of all four approaches can be summarized in a bar chart as shown in Fig. 5.

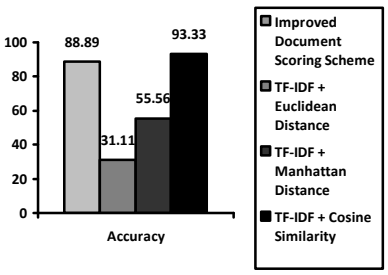

Fig. 5. Classification accuracy of different appraoches

V. CONCLUSION

In this research work, our main target was to create a vector space model for retrieving topic of a Bengali document. Vector space model is a traditional similarity based retrieval model which is a matrix with documents placed in rows and all vocabulary terms placed in columns. In this work, the weight values of this model was calculated using TF-IDF weighting and Improved Document Scoring Scheme weighting techniques. Three similarity measures namely Euclidean distance, Manhattan distance and Cosine Similarity measure were used with the TF-IDF weighting scheme. Finally the best retrieval accuracy was obtained using TF-IDF weighting scheme with cosine similarity measure, which is 93.33%. Our future work encompasses a target to increase the retrieval accuracy and employ other retrieval techniques as well.

REFERENCES

[1] K. Sparck Jones and P.Willett, editors. Readings in Information Retrieval. Morgan Kaufmann Publishers, 1997.

[2] T. Hofmann, "Probabilistic latent semantic analysis". In Proceedings of UAI 1999, 1999, pp. 289–296.

[3] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation". Journal of Machine Learning Research, vol. 3, 2003, pp. 993–1022.

[4] Al-Shargabi B., Al-Romimah W., Olayah F., "A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination", in Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, 2011, pp. 11:1-11:5.

[5] Hmeidi I., Al-Ayyoub M., Abdulla N.A., Almodawar A. A., Abooraig R., Mahyoub N.A., "Automatic Arabic Text Categorization: A Comprehensive Comparative Study", Journal of Information Science, vol. 41, issue 1, 2015, pp. 114-125.

[6] Al-Shalabi R., Obeidat R., "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing", INFOS2008, 2008, pp. NLP108-NLP112.

[7] Rajan K., Ramalingam V., Ganesan M., Palanivel S., and Palaniappan B., "Automatic classification of Tamil documents using vector space model and artificial neural network," Expert system with Applications, vol 36, 2009, pp. 10914-10918.

[8] Jain U., Saini K., "Punjabi Text Classification using Naive Bayes Algorithm", International Journal of Current Engineering and Technology, vol. 5, issue 6, 2015, pp. 3777-3779.

[9] Liliana D.Y., Hardianto A., Ridok M., "Indonesian News Classification using Support Vector Machine", International Journal of Computer and Information Engineering, vol. 5, No. 9, 2011, pp. 1015-1018.

[10] Islam M.S., Jubayer F.E.M., Ahmed S.I., "A support vector mixed with TF-IDF Algorithm to categorize Bengali Document", in proceedings of Internation Conference of Electrical, Computer and Communication Engineering, 2017, pp. 191-196.

[11] Mandal A. K. and Sen R., "Supervised learning methods for bangla web document categorization", International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 5, No. 5, 2014, pp. 93-105.

[12] Kabir F., Siddiquey S., Kotwal M.R.A., Huda M.N., "Bangla Text Document Categorization Using Stochastic Gradient Descent (SGD) Classifier", International Conference on Cognitive Computing and Information Processing, 2015, pp. 1-4.

[13] Chy A.N., Seddiqui M.H., Das S., "Bangla News Classification using Naive Bayes classifier", in proceedings of 16th International Conference on Computer and Information Technology, 2014, pp. 366-371.

[14] Mansur M., UzZaman N., Khan M., "Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus", Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh.

[15] Amit Singhal. "Modern information retrieval:A brief overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 24, issue 4, 2001, pp. 35–43.