Myanmar News Retrieval in Vector Space Model using Cosine Similarity Measure

Hay Man Oo
Natural language Processing Lab
University of Computer Studies
Yangon
haymanoo@ucsy.edu.mm

Win Pa Pa
Natural language Processing Lab
University of Computer Studies
Yangon
winpapa@ucsy.edu.mm

Abstract

Information Retrieval (IR) is an effective means of retrieving the best relevant document to user query. Nowadays, the problem of documents similarity deals with IR is retrieving required information from a large amount of data. In this paper we studied a Vector Space Model (VSM) that is used in IR and represents a document as a vector in an n-dimensional space, where each dimensional represents a term and measured through cosine angle between two vectors. The objective of this paper is to retrieve the relevant file from Myanmar news data sets using cosine similarity measure in VSM to user's query. Evaluations are done in terms of similarity score by Precision, Recall and F-score.

Keywords— *Information Retrieval, TF-IDF weighting scheme, cosine similarity measure, Vector Space Model*

I. INTRODUCTION

The basic form of Information Retrieval (IR) is based on an input query for retrieving of documents. Nowadays, many users are searched documents on the web. The complete example for this application is web search. Web search engines are the most visible in IR applications. For this purpose, developed many algorithms accept a user query and search it in the documents collection and rank the results of similarity score relevant to the user query. Against the individual query terms, these algorithms based on maintain the information deals with term frequencies and positions which matched to indexed documents. Assigning to each document focused on its value is similarity score. High for a high frequency in the document is the query term's score. In analyzing this similarity and computing the score, different algorithms take different approaches. To do this work, the Vector Space Model (VSM) is one of the best approaches. It overcomes the Boolean Model, which uses Boolean desired queries based on Boolean logic and searches it in the documents, retrieval result is based on either the desired terms are in the documents or not. This gives too many or too few documents [9]. The role of textual information retrieval is term weighting. Term Frequency and Inverse Document Frequency (TF/IDF) weighting scheme is one of the most popular schemes of term weighting, followed by Okapi BM25. We discuss three modified schemes in the following text. A new term weighting scheme, Term Frequency with Average Term Occurrences (TF-ATO), [3] which is an advance than the TF/IDF weighting scheme. These schemes use a discriminative approach depend on the document based vector to discard fewer significant weights from the documents and calculate the average term occurrences of terms in documents. TF-ATO effect of discriminative approach and the stop words removal process on the IR system capability and achievement than using the well-known TF-IDF that is an importance part in the Vector Space Model (VSM). Another modified TF-IDF scheme [4] which exploits two features of within document term frequency normalization to decide the critical of a term. One component of the TF tends to adopt short documents, while the other tends to adopt long documents and combine these two TF components using the query length information that keeps a balanced trade-off in retrieving short and long documents, when the ranking function allows queries of different lengths. The IDF of a term is similar distance between the empty string and the term which is approximated shown that M. Shirakawa [5]. They have proposed a global term weighting technique, N-gram IDF, by collaborating IDF and described the clarity and durability of N-gram IDF on key term abstraction and Web search query tokenization functions. It able to achieve competitive performance with advanced methods planned for any task adding efforts and assets.

II. MYANMAR INFORAMTION RETRIEVAL

One of the Natural Language Processing (NLP) advanced techniques, Information retrieval (IR) is defined to be the science of enhancing the effectiveness of term-based document retrieval. An IR system is an information system that collects, indexes and keeps the data for extracting of relevant information responding to a user's query (user information need). An IR process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. IR is searching material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large. In this paper, we used Myanmar information retrieval for retrieving file containing Myanmar news documents from Myanmar news corpus that is relevant to user query.

III. DATA COLLECTION

In this part, we collects Myanmar news corpus about 7000 documents written with Myanmar Unicode font from Mizzima Burmese news website. Types of news are shown in Table 1.

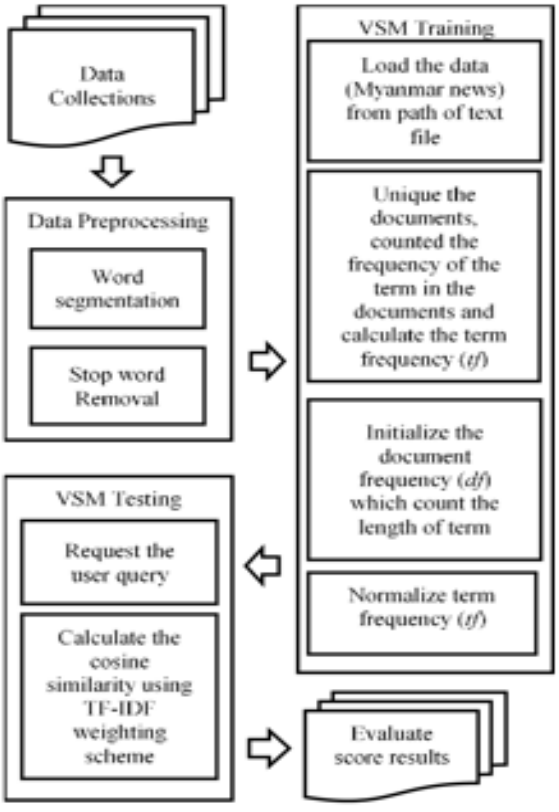

Figure 1. The process of Myanmar news retrieval

TABLE I. STATISTIC OF MYANMAR NEWS CORPUS

Type of news	Number of documents	Number of documents
Health	2115	18195
Sport	1349	2865
Entertainment	676	6277
Polictical	1423	15538
Economic	1437	11836
Total	7000	54711

IV. MYANMAR NEWS RETRIEVAL

Myanmar online news was collected for News Corpus from Myanmar news website and was done the word segmentation and stop words removal as in data preprocessing steps. Vector Space Model (VSM) using cosine similarity measure is applied in this work to train and test the corpus for the final output is the similarity score that are relevant to user query, and the process is shown in Fig. 1. We used about 700 corpus in training phrase and about 4000 corpus in testing phase. Preprocessing steps is described in section V with detail description of training phase with four modules: loading the Myanmar news data collection, unique the documents in data collection and counting the term *t* occurrences in the documents *d* and calculation the term frequency (*tf*), Initializing the document frequency (*df*) that is the document *d* occurrences in the term *t* and the last is normalized term frequency (*tf).* Testing phase has contained two parts: one is requested the user query and tokenized these query and then calculated the cosine similarity measure using TF-IDF weighting scheme and the final results is relevant to user query.

V. DATA PREPROCESSING

In this step, we collected Myanmar news data from Mizzima Burmese news website. All of this news is used Zawgyi fonts. Therefore, we transformed to Unicode fonts, tokenized this news and removed the stop words as described below [13].

A. Word segmentation

Word segmentation is the fundamental task in natural language processing. This process split into word and sentences that are searching word tokens and borders of sentences. English word boundaries can be

easily defined but it is not in Myanmar. In Myanmar word boundaries, words are frequently written without spacing in sentences. Therefore, for pass sentences and word tokens, word segmentation is very useful for Information Retrieval (IR). For this purpose, we used Myanmar word segmenter of UCSY [10]. In preprocessing steps, word segmentation is very important for evaluation of IR.

B. Stop word removal

The objective of stop word removal is filter out words that appear in most of the documents. The examples of stop words removal are also shown in Table 2 [12]. This step is very important step in preprocessing techniques used in Natural Language Processing application.

TABLE II. EXAMPLE OF PREPROCESSING SENTENCES

Original sentences	Preprocessed sentences
ဧရာဝတီတိုင်း တွင် တွေ့ ရှိခဲ့ သော ဒုက္ခသည် များ ကို ရခိုင်ပြည်နယ် တွင် ယာယီ ထား‌ရှိ ရေး	ဧရာဝတီတိုင်း
	တွေ့ ရှိ
	ဒုက္ခသည်
	ရခိုင်ပြည်နယ်
	ယာယီ ထား‌ရှိ ရေး
အမျိုးသား လူ့ အခွင့်အရေး ကော်မရှင် ထံ တစ်ပတ် အတွင်း တိုင်ကြား သွား‌မည်	အမျိုးသား လူ့ အခွင့်အရေး ကော်မရှင်
	တစ်ပတ်
	တိုင်ကြား
	သွား
ရွေးကောက်ပွဲ ကော်မရှင် ၏ လုပ်ဆောင်ချက် များ သည် ဥပဒေ နှင့် မ ညီ သည့် အမှား များ အပြင် မ မှား သင့် သည် များ ကိုပါ မှားယွင်း နေသည် ဟု ဒေါ်အောင်ဆန်းစုကြည် ဦးဆောင်သည့် အမျိုးသား ဒီမိုကရေစီ အဖွဲ့ချုပ် (NLD) က ဖွန်လ ၃ ရက် တွင် အိတ် ဖွင့် ပေးစာ တစ်စောင် ပေးပို လိုက် သည် ။	ရွေးကောက်ပွဲ ကော်မရှင်
	လုပ်ဆောင်ချက်
	ဥပဒေ
	မ ညီ
	အမှား
	မ မှား သင့်
	မှားယွင်း
	အောင်ဆန်းစုကြည်
	ဦးဆောင်
	အမျိုးသား ဒီမိုကရေစီ အဖွဲ့ချုပ်
	NLD
	ဖွန်လ ၃ ရက်
	အိတ် ဖွင့် ပေးစာ
	တစ်စောင်
	ပေးပို
ရှမ်းပြည်နယ် တောင်ပိုင်း ပြည်သူ ဖိရမ် တွင် ဖက် ဒရယ် မူ နှင့်ပတ်သက်ပြီး ဒေသခံ ပြည်သူများ နှင့် အရပ်ဘက် အဖွဲ့အစည်းများ သေချာစွာ သိရှိ နားလည် စေရန် ဆွေးနွေးပွဲ တစ်ခု ထည့်သွင်း ဆွေးနွေး သွား‌မည်	ရှမ်းပြည်နယ် တောင်ပိုင်း ပြည်သူ ဖိရမ်
	ဖက် ဒရယ် မူ
	ပတ်သက်
	ဒေသခံ ပြည်သူ
	အရပ်ဘက် အဖွဲ့အစည်း
	သေချာ
	သိရှိ နားလည်
	ဆွေးနွေးပွဲ တစ်ခု ထည့်သွင်း ဆွေးနွေးသွား
	ယခုနှစ် ကုန် ပိုင်း
ယခုနှစ် ကုန် ပိုင်း အတွင်း ကျင်းပ ရန် လျာ ထား‌သည့် အထွေထွေရွေးကောက်ပွဲ တွင် အစိုးရအဖွဲ့ က အဖွဲ့အစည်း အသီးသီး နှင့် ပူးပေါင်း မည်	ကျင်းပ
	လျာ ထား
	အထွေထွေရွေးကောက်ပွဲ
	အစိုးရအဖွဲ့
	အဖွဲ့အစည်း အသီးသီး
	ပူးပေါင်း

VI. VECTOR SPACE MODEL

The Vector Space Model (VSM) represents documents and queries as vectors in multidimensional space, whose dimensions are terms used to build an index to represent the documents. In information retrieval used the VSM, indexing and relevancy rankings and can be successfully used in evaluation of web search engines. The VSM procedure can be divided into three steps. The first is the document indexing where content carrying terms are extracted from the document. The second is the weighting of the indexed terms to improve retrieval of documents relevant to the user. The last step ranks the document with respect to the query according to a similarity measure. A common similarity measure known as cosine determines angle between the query vector and the document vector as described in next section. The angle between two vectors is considered as a measure of difference between the vectors, cosine angle is used to calculate the numeric similarity, determines angle between the query vector and the document vector. The VSM is an algebra model for characterizing text documents which outperform the Boolean Model limitations. The major advantage of VSM is used the weights applied to the term which not binary. VSM grants for more effectively eliminating results and evaluating over similarity range values than the Boolean Model. This model means vectors of weights as documents and queries. Each weight is a measure of the important of an index term in a document or a query, commonly. The index term weights are applied on the basic of the density of the index terms in the document, the query or the collection. As a further process, the document can be decreases different word forms into a useful stem which provides improve the matching the similarity between each documents. As to give better results, the query terms perhaps weighted assign to their related importance. This paper tends to relevance the file in documents from the user query.

A. Cosine Similarity Measure

In Vector Space Model (VSM), the query and documents are represented by a two-dimensional vector. The vector cosine measure is a useful method

to measure the similarity between the two vectors. It is measured by the cosine of the angle between the two vectors and determines whether two vectors are pointing in the same direction (Fig. 2). For considerate this similarity between the two vectors, Euclidean distance is a bad measure and it does not give for realistic results. The two vectors are at 90 degrees to each other is 0 for the cosine value and it has no match. The smaller the angle and the greater the relevant between vectors is near the cosine value to 1. If the documents are similar, evaluation of 0^o angle is 1 of similarity score and if the documents are also entirely dissimilar, evaluation of angle 90^o is 0 of similarity score.

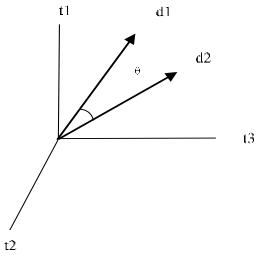

Figure 2. Illustration of angle similarity between two documents

For comparing their weights, implementation of length normalized vectors is the cosine weighting measure. For Cosine Similarity gives the formula, as in (1).

$$Similarity = cos(d', q') = \frac{d'.q'}{||d'||.||q'||} = \frac{\sum_{i=1}^{|V|} d_i q_i}{\sqrt{\sum_{i=1}^{|V|} d_i^2} \sqrt{\sum_{i=1}^{|V|} q_i^2}} \quad (1)$$

B. Term Weighting Scheme

The term frequency, $tf_{t,d}$ is times of the term t appears in document d. The term t computes the document-query match scores. The log-frequency weighting with term in the documents is defined as,

$$W_{t,d} = 1 + log_{10}(tf_{t,d}) \quad (2)$$

Rare in the collection for a term in a query that is a document. To apply for this, document frequency, df_t is the document *d* occurrences appear anywhere occurs the term *t*. idf_t means an inverse document frequency is also usefulness of the term *t*.

$$\text{idf}_t = \log_{10}(\frac{N}{df_t}) \quad (3)$$

The *tf-idf* is the product together with *tf* weight and *idf* weight that is one of the perfect weighting schemes in the information retrieval.

$$W_{t,d} = (1 + log_{10}(tf_{t,d})) * log_{10}(\frac{N}{df_t})) \quad (4)$$

It increases in a term occurrences appearing a document and the term query rarity in the documents collection.

VII. Experiments and Experimental Results

This paper evaluated Myanmar news data sets over 7000 documents and used Vector Space Model (VSM) with cosine similarity measure. As our experiments, we used the VSM python module to retrieve documents automatically. The similarity score results retrieved from Myanmar news documents corpus that is relevant to user query.

As our experimental results, the score of similarity results are relevant to user query. In preprocessing steps, we initially collects Myanmar news corpus about 7000 documents from Mizzima websites and used Myanmar word segmenter of UCSY [10] for Myanmar sentences and word segmentation. In our experimental results, we prepared Myanmar words from collected Myanmar news corpus for user query to request their required information. As in Fig. 3, we randomly used one to nine words user queries about 300 words search, T1-T3 described one to three words about 100 words search, T4-T6 also described four to six words and T7-T9 also described seven to nine words about 200 words search. The Fig. 3 shows the average Precision, Recall and F-score value obtained by calculating Vector Space Model (VSM) for 300 user queries. The VSM rank the documents with their similarity score. Each score calculated by VSM measured with cosine similarity. The documents having zero scores are irrelevant and assigned the lowest possible rank. Therefore, their ranks are required for calculating Precision, Recall and F-score. These values are used in measure the relevancy of documents. The results of Fig. 3 for retrieving the top documents as our threshold value and calculated by Precision, Recall and F-score. As shown in Fig. 3, the maximum and minimum of Myanmar words queries are present in the numbers of words length.

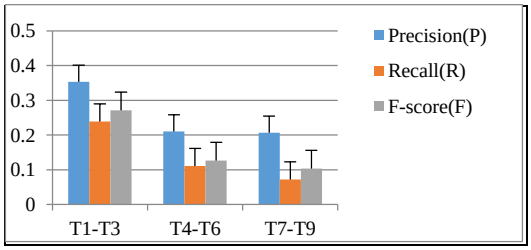

Figure 3. Similarity results used cosine in vector space model for user query

The maximum numbers of T1-T3 means one to three words search queries and the minimum numbers of T7-T9 also means seven to nine words search queries. As this result, we studied that the ranking of documents can vary with words search query length because of their weighting scheme. So, T1-T3 is highest rate in other query length and T7-T9 is also lowest rate in other query length. The final results are obtained the similarity score of T1-T3, T4-T6 and T7-T9 in Precision, Recall and F-score respectively shown in Fig 3.

VIII. CONCLUSION AND FUTURE WORK

In this paper analyzed approach of Vector Space Model (VSM) for check retrieval queries. The similarity value is calculated by using approach of VSM. After analyzing the weighting terms in document collection, was evaluated by similarity value between queries and documents. Documents ranking based on the score of similarity value evaluated by VSM approach. Experiments applied on Myanmar news data sets and the proposed model show that outperforms relevant file in Myanmar news data sets to user query. In this field, future work would be developing new similarity measures, new weighting schemes and new models that can efficiently focused on a huge amount of data sets utilizing semantic information.

REFERENCES

[1] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

[2] Singh, Vaibhav Kant, and Vinay Kumar Singh. "Vector space model: an information retrieval system" Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March 141 (2015): 143.

[3] Ibrahim, O., and D. Landa-Silva. "Term frequency with average term occurrences for textual information retrieval." Soft Comput 20.8 (2016): 3045-3061.

[4] Paik, Jiaul H. "A novel TF-IDF weighting scheme for effective ranking." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.

[5] Shirakawa, Masumi, Takahiro Hara, and Shojiro Nishio. “N-gram IDF: a global term weighting scheme based on information distance” Proceedings of the 24th International Conference on World Wide Web. ACM, 2015.

[6] Choi, Seung-Seok, Sung-Hyuk Cha, and Charles C. Tappert. "A survey of binary similarity and distance measures." Journal of Systemics, Cybernetics and Informatics 8.1 (2010): 43-48.

[7] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In Proceedings of Text Retrieval Conference (TREC), pages 109–126, 1994.

[8] Robertson, Stephen, and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc, 2009.

[9] Arash Habibi Lashkari and Feresteh Mahdavi, “A boolean model in information retrieval for search engines”, 2009 International Conference on Information Management and Engineering.

[10] Win Pa Pa and Ni Lar Thein. 2008. “Myanmar word segmentation using hybrid approach.” In Proc. of ICCA. 166–170.

[11] Joydip Datta , Dr. Pushpak Bhattacharyya. “Ranking in information retrieval”, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay Powai, Mumbai – 400076.

[12] Htay, Hla Hla, G. Bharadwaja Kumar, and Kavi Narayana Murthy. “Statistical analyses of Myanmar corpora” Technical report, Department of Computer and Information Sciences, University of Hyderabad. 26 march, 2007.

[13] Nyein Thwet Thwet Aung and Ni Lar Thein . “Word sense disambiguation system for Myanmar word in support of Myanmar-English machine translation” University of Computer Studies, Yangon, Myanmar, 2011.