ISSN 1512-1232

SEMANTIC TEXT CLUSTERING USING ENHANCED VECTOR SPACE MODEL USING NEPALI LANGUAGE

Chiranjibi Sitaula

Central Department of Computer Science and Information Technology,Tribhuvan University,Kirtipur,Kathmandu,Nepal.
candsbro@gmail.com

Abstract

We propose an algorithm which combines the advantage of classical vector space model to cluster the semantic texts.Those text having similar context words are taken as the semantic texts. So as to remove the deficiencies of the classical vector space model,which was not able to cluster such text, the concept of advanced of enhanced vector space model is proposed.It takes the concept of fuzzy set theory.The enhanced vector is obtained by adding the ***tf-idf*** *with fuzzy membership value and perform the cosine operation in order to calculate the semantic distance between the text.*

Keywords*: VSM(vector space model),Clustering,Fuzzy set.*

1. INTRODUCTION

Clusteing is the process of grouping the texts depending on the distance between them.The distance is defined as the correlation between the texts which may be cosine or euclidean distance.Semantic texts are those texts having similar meanings in particular context.If the groups of those texts are found, then it is termed as the semantic text clustering.

VSM is the model where each term represents the number of dimension of the documents present in the particular document.The VSM is riched by different terminologies like term-frequery, inverse document frequency and cosine similarity.

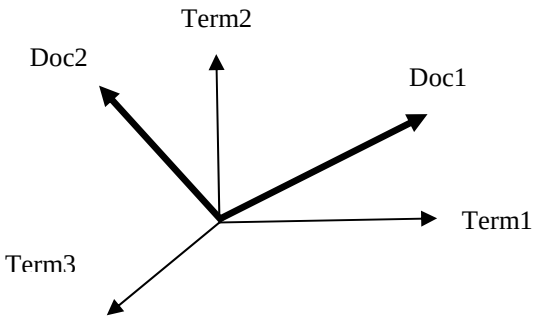

Figure 1. Vector Space Model

The fundamental terms of the VSM is defined as follows:

TF-Term Frequency,IDF-Inverse Document Frequency. The formula for the vector space model as given by[1] is as follows

*Tf-idf=tf*log(D/dfi) where, D=no.of documents,dfi=no.of documents having that particula r term and idf=log(D/dfi).*

According to [2], the Gaussian membership function is given by the following formula in the exponential form

$$\mu_{FNS}(z) = \exp(-(z - \bar{z})^2 / \sigma)$$

where z is the term, $\bar{z}$ is the mean of the gaussian membership function and σ determines the size

ISSN 1512-1232

of the gaussian shape.

We have used the cosine similarity measure for calculating the similarity between two documents A and B as defined by [2] is

$$\cos(\theta) = \frac{A.B}{|A|.|B|}$$

According to [8], fuzzy set theory is used for the artificial intelligence part.It is an extension of the classical set theory because in the classical set theory only true or false were taken as the decision case but in this model, beyond true or false decision, the degree of truthness is calculated using different strategies availble in this model. The degree of truthness is calculated using like Gaussian membership function or trapezoidal,triangular etc.

The paper is divided into following sections:INTRODUCTION contains information about the research, LITERATURE REVIEW contains the information about the existing research into this topic,PROPOSED METHOD contains information about proposed algorithms for text clustering and CONCLUSION contains the final result of the proposed algorithms.

2. LITERATURE REVIEW

The clustering task has been performed by different researchers. [1] took the concept of modified vector space model in which they modified the inverse document frequency with document frequency only. Similarly, [2] took the vector space model for blog analysis purpose. They used fuzzy based method for the blog and its respective contents. [3] used vector space model for the clustering purpose and they used the k-nearest neighborhood method for the clustering problem.[4] used vector space model for text categorization process. They used it for enhancing nearest neighborhood method. [5] used corpus based concept for the word similarity. They used context-vector space model for the similarity of the words. [6] used graphical method with **WORDNET**(an online dictionary) for clustering of blogs with enhanced semantics. They found that the graph based model performance was better and they did not focus much about Vector Space Model. [9] discussed about the unsupervised approach. i.e. **Point wise Mutual Information** (PMI) for finding the semantic relevancy of the text using search engine. [10] used dynamic programming approach and **PMI** for the semantic text similarity.

3. PROPOSED ALGORITHMS

Clustering the text using the normal vector space model could not handle the semantic relevancy of words so due to lack of such features in traditional vector space model the concept of enhanced vector method is proposed. The research has not been performed yet in opinion mining task in Nepal which is the leading task for Nepali researcher who wants to work in Nepali language for the text clustering. The algorithms which work in English language may not work in other language. The clustering task enables the analyst to observe those clusters having maximum number of documents which saves the time in this busy world for the opinion to be analyzed by the analyst.

The complete algorithms of proposed model is given in Figure 2 (see below).

The algorithms given in Figure 2 is the complete model for the proposed semantic text clustering using Nepali Language.In this approach, first step is accompanied by the calculation of the term frequency and inverse document frequency which is followed by the multiplication of the term frequency and inverse document frequency for each term.After calculating the product of *tf*idf*, it is followed by calculation of the membership value from the fuzzy for each individual keyword which is given in the set.If the set doesnot contain then the membership value 0 will be added to the *tf*idf* value, otherwise the value obtained after calculation will be added.Also, the query vector is responsible for maintaining the semantic meanings. While preparing the query vector, if the same fuzzy set contains the terms present in one document then the membership value will be added to each of the value which either may be 0 or 1 in the query vector. In this way, the algorithm is implemented.

Every document is represented by query vector and document vector. The query vector is

ISSN 1512-1232

applied to each individual document vector for checking the cosine similarity. In this research, while implementing in the single keyword document, the restrictions is not made but in the multi word document, the rigid format of the sentence is given.

Step1: Calculate the term frequency and inverse document frequency of the document keywords.
Step2: Calculate the *tf***idf* value of each keyword.
Step3: Calculate the value of membership of the term from fuzzy set.
Step4: New vector value=*tf***idf*+membership value.
Step5: Calculate the cosine similarity of the documents for clustering.
Step 6: If the cosine value>0.50 Then cluster that into one cluster
Else goto Step5 and perform the same operation with another document vectors.
Step7:Obain the semantically related clusters.
Step8:End.

Figure 2. An algorithms for the semantic text clustering

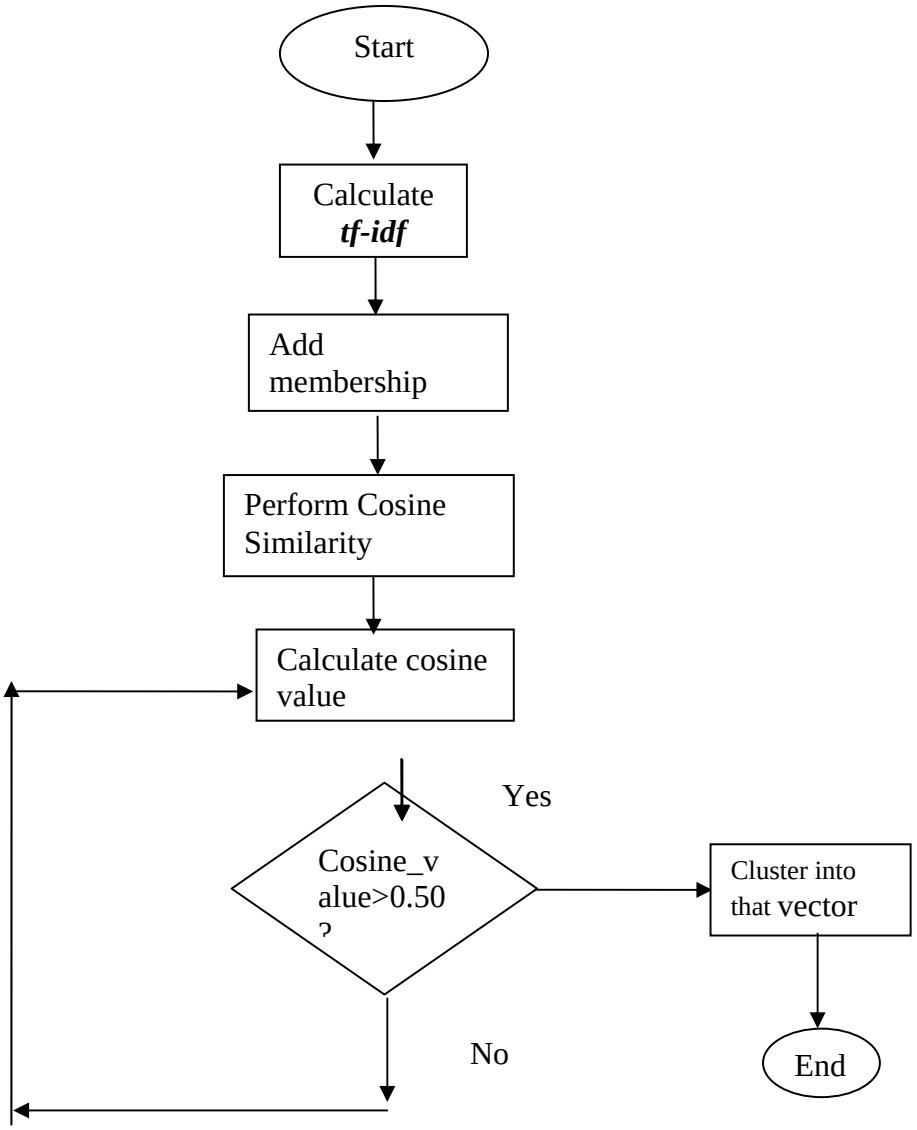

Figure 3. Flow Diagram for the semantic text clustering.

Figure 3 is the diagram for showing how the operations are flowed.Through this diagram, it can shown that how the algorithms actually works in real scenario. In the proposed model, for the

ISSN 1512-1232

single keyword document it works properly. But in the multi keyword document level the single format of data is taken.

4. EXPERIMENTAL RESULT

The experimental result of this research was conducted in two ways. Firstly, single keyword documents was taken and second the multi-keyword documents were taken. The performance of the cluster were analyzed by using random index as explained by[7]. The research was implemented in *php* language. Inputs were given using file system. They were kept in the files.

In single word document, the result obtained for two different set of data set which were in Nepali Language

The result obtained from the experiment is shown in Table 1 and Table 2 below.

For single word document

No of documents	Random Index	Accuracy
45	0	0.90

Table 1. Experimental result for single keyword

Similarly, for the multi word document

No of documents	Random Index	Accuracy
45	0	0.91

Table 2. Experimental result for multi keyword document

Some of the snapshots of the outputs are given in Figure 4 and Figure 5 as follow.

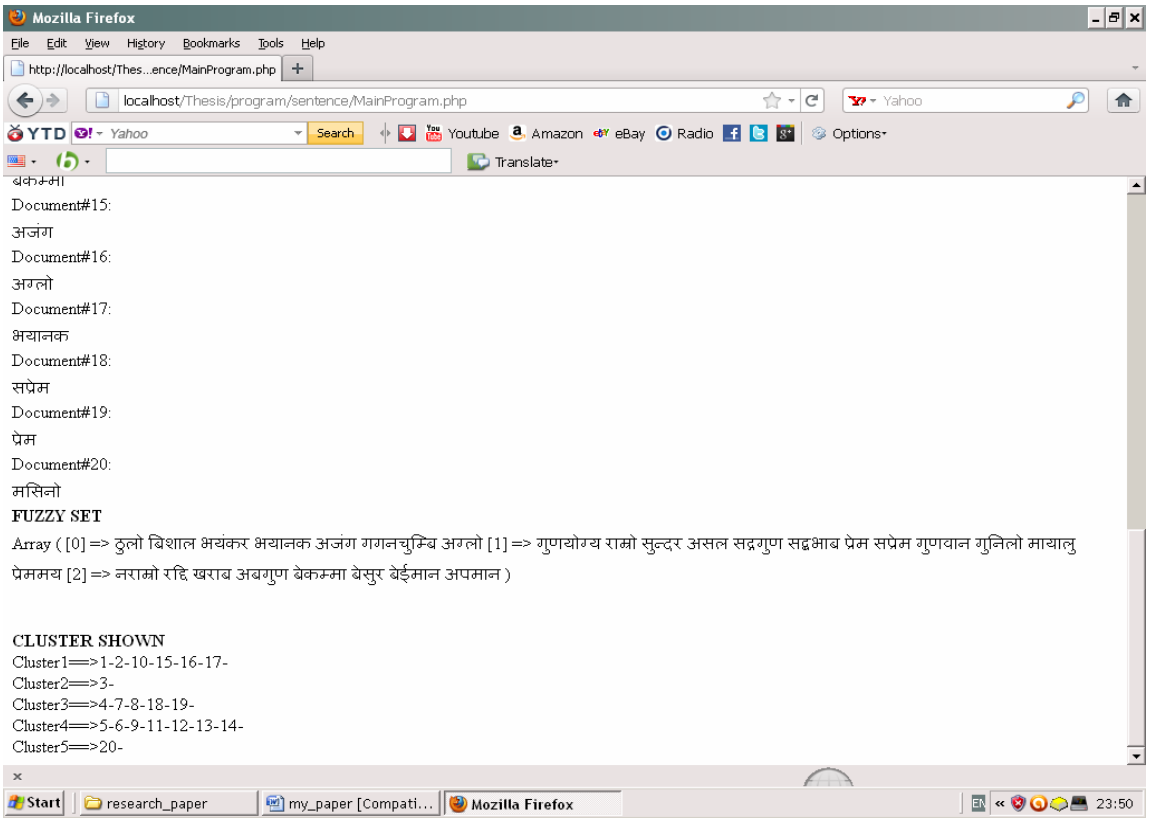

Figure 4. Output obtained for single keyword document

Figure 4 is the output obtained from the documents taken as the single keyword documents. This output showed that there were no any intersection of the documents so that the value of the

ISSN 1512-1232

random index is influenced. But random index remained zero in this experiment.

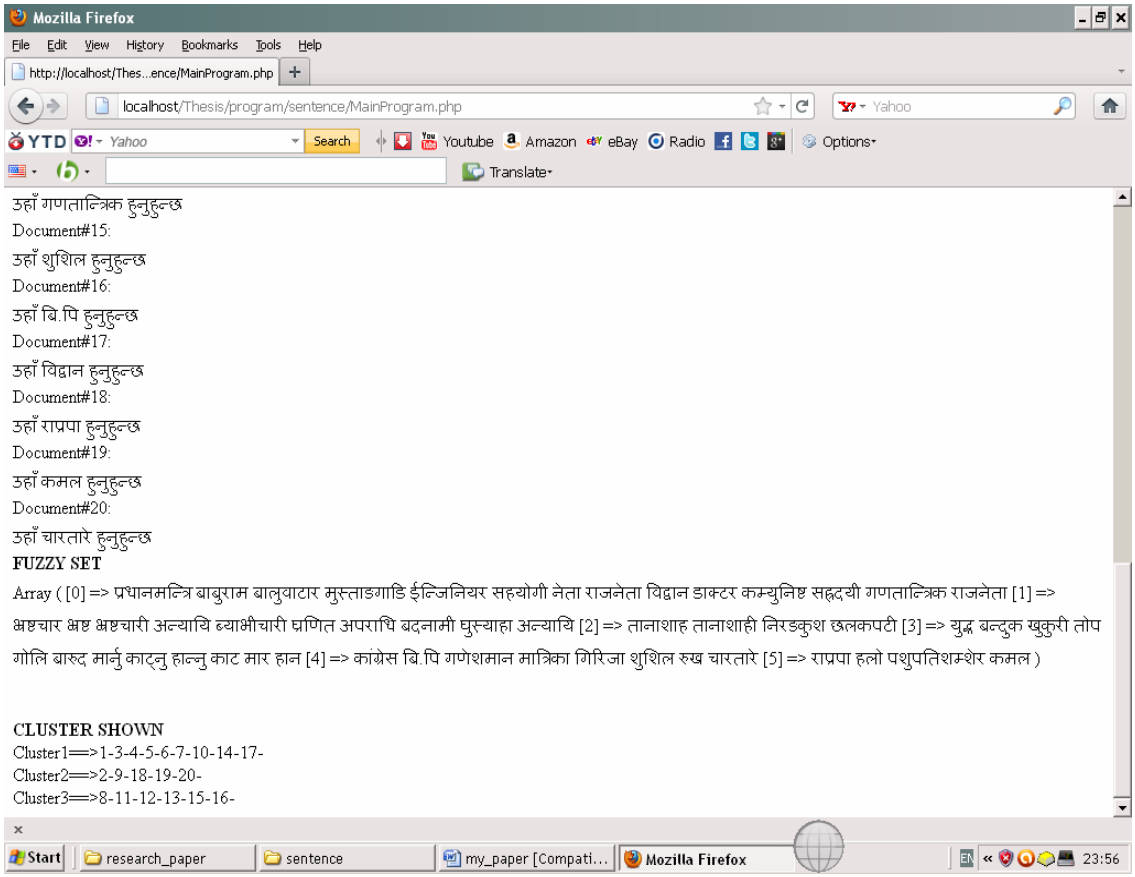

Figure 5. Output obtained for multi word document

The output obtained using multi keyword is shown in the figure 5. It has been shown that the algorithms is working properly since it has no intersection of the documents that affects the random index.

5. CONCLUSION AND FUTURE WORKS

In this work semantically similar documents in both word level and the document level are selected completely using unsupervised methods.

Since this thesis takes keywords in each document to represent the documents and then use these keywords to identify document similarities. Classical ***tf-idf*** couldnot not distinguish the semantic words although it was used to keyword's importance. By applying fuzzy semantic to the Vector Space Model was able to explore the hidden relationship between documents.

Due to the lack of such features in classical vector space model, the concept of enhanced Vector Space Model is proposed which produced promising result.

In this methodology, the classical vectors are enriched with adding the fuzzy membership values from the fuzzy set made. After enriching the vectors with such values, it becomes ready for normal vector operation which we used in linear algebra. The classical vector is based on ***tf-idf*** approach and also in this thesis same concept is used but ***tf-idf*** is not complete vector for this thesis because after calculating the weight (***tf-idf***), degree of truth ness value i.e. membership values should be added for the respective term with the corresponding terms in the document vectors.

A large number of experiments have been made (About 100^{th} Nepali documents in both word and sentence level). The experimental results are analyzed using Random Index and number of luster produced.

ISSN 1512-1232

It can be concluded from this study is that fuzzy set and vector space model, when combined together, performed semantic text clustering easily. But for this, it would be better if the number of fuzzy set and number of keywords is increased.

This new approach has promising result and the result can further be improved with the following future work:

- The large number of fuzzy set can be used. This can be enriched with more semantically related words for each set.
- The number of testing documents can be increased to check its consistency and performance with lots of variants.
- The concept of stop words was not taken into consideration for document processing because of Nepali Unicode which is under research itself.
- The Gaussian membership function is used in this thesis for calculating the degree of truth ness from the fuzzy set. In place of Gaussian membership function, others function like trapezoidal, triangular can be used to check whether it works properly or not.
- The concept can be applied to paragraph level and documents having multiple paragraphs.

References

1. Abdul-Rub, Mohammed Said,"A modified vector spaced model for protein Retrieval", UHCSNS, Vol 7 No 9, 2007.
2. Ho, Chi-Shu,"Blog analysis with Fuzzy TFIDF", Master Project, San Jose State University, 2007.
3. Jaiswal, Mayank Prakash, "Clustering Blog Information", Master Project, San Jose University,Paper 36, 2007.
4. Shin Kwangcheol , Abraham Ajith, Han Sang Yong, " Improving kNN Text Categorization by Removing Outliers from Training Set",Springer-Verlag Berlin Heidelberg, pp. 563-566, 2006.
5. Emre Esin Yunus, "Improvement of corpus-based word similarity using vector space model", Mater Thesis, Middle Ease University, 2009.
6. Singh, A.K, Joshi, R.C, "Clustering of Blogs with Enhanced Semantics", International Journal of Computer Applications (0975-8887), Volume 16-No.7, February 2011.
7. Perumal, P, Nedunchezhian, "Performance Evaluation of Three Model-Based Documents Clustering Algorithms", European Journal of Scientific Research VOL. 52-No.4, 2011.
8. Chakraborty, RC. "Fuzzy Set Theory," 2010.Accessed Date: 19th July, 2011.
9. Liu, Bing, "Sentiment Analysis and Subjectivity", *Handbook of Natural Language Processing*, Second Edition, 2010.
10. Islam, Aminul, inkpen, Diana, "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity", ACM Transaction on Knowledge Discovery from Data, VOL. 2, No.2 , Article 10,, July 2008

Article received: 2011-11-16