Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 6; October-December, 2016, pp. 512-516

http://www.krishisanskriti.org/Publication.html

A Comparative Analysis of K-means, Particle Swarm Optimization and Self- Organizing Map for Clustering of Sense Tagged Nepali Documents

Sunita Sarkar[1] and Arindam Roy[2]

[1,2]Department of Computer Science Assam University Silchar
E-mail: [1]sunitasarkar@rediffmail.com, [2]arindam_roy74@rediffmail.com

Abstract—*Human Languages are ambiguous. There is possibility of different meanings of a particular word in different context. So consideration of senses of words is very important for representation of documents in document clustering. The common way to represent a text is bag of words approach. Words/ terms that are present in the document are considered for generation of feature vector. The feature vector representing the document is generated by using the frequency count of terms in a document. Vectors so generated by counting the frequency of terms cannot properly represent synonymy and polysemy. In this paper feature vectors are generated using sense tagged documents. All the terms/ words in the sense tagged documents provide the exact senses of the words. Information about the senses of the words is obtained from Wordnet. WordNet is used as the sense inventory for tagging of words in the documents . The feature vectors generated from sense tagged documents consider synset ids as element of the vector rather than terms. Standard K-means algorithm, Particle Swarm Optimization(PSO), hybrid Particle Swarm Optimization-Kmeans and hybrid Self Organizing Map-Kmeans algorithms are applied for clustering of documents in Nepali language. Experimental results show that when feature vectors are generated from sense tagged documents clustering tends to perform better than when the term based method is used.*

Keywords*: Sense tagged document, Nepali WordNet, Synset id;*

1. INTRODUCTION

Due to advances in information technology, cloud computing and internet of things etc. data has been increasing at exponential rate in recent years. The data in the web may be textual data, image data , video and so on. The main challenge is to extract useful information and knowledge from the large amount of data available in the web. There are many Search engines which are powerful information retrieval tool aid in searching textual data or documents by category, content, or topic. Clustering[1] or classification techniques are applied in order to make such information retrieval more effective. An essential step for processing all these applications is to represent documents in a format suitable for these applications. The traditional way to represent documents are as vectors and is the most accepted and popular method. The feature vectors representing the document are constructed by using the frequency count of terms in a document. These methods treat documents as a bag of independent words. One of the limitations of these methods is that they are unable to properly handle linguistic phenomena such as polysemy where a word has different meanings in different context (eg. "fly" as fly in the air vs. "fly" as an insect) or synonymy where different words have same meanings (eg. search "taxi" when the user queries for “cabs”). Thus inherent meaning of the content of a document can 't be captured by the term based methods. In this work an effort has been made to represent the document with their senses. Sense information of the words is obtained from Wordnet. WordNet is used as the sense inventory for sense tagging of documents. Sense tagging is the task of identification and tagging a particular sense to the word in the given context out of many senses available for that word. In this work text documents used are in Nepali[1] Language. K-means, Particle Swarm Optimization(PSO), hybrid Particle Swarm Optimization-Kmeans and hybrid Self Organizing Map-Kmeans algorithms clustering algorithms are applied to the set of vectors generated using Nepali sense tagged documents for clustering. The rest of the paper is organized as the following.. Section 2 provides an overview of WordNet. Section 3 discusses about sense tagged corpus. Section 4 describes about construction of document vectors using sense tagged corpus. Clustering algorithms are presented in section 5 Experimental results are reported in section 6 . The paper is concluded in section 7.

2. WORDNET

WordNet is a machine readable lexical database whose design is inspired by current psycholinguistic theories of human lexical memory. In WordNet lexical information are organized

[1]Nepali is an Indo-Aryan language spoken by approximately 45 million people in Nepal, where it is the language of government and the medium of much education, and also in neighboring countries (India, Bhutan and Myanmar). Nepali is written in the Devanagari alphabet. It is written phonetically, that is, the sounds correspond almost exactly to the written letters. Nepali has many loanwords from

Arabic and Persian languages, as well as some Hindi and English borrowings [3].

in terms of word meanings/ concepts, rather than word forms[4]. Princeton English WordNet [5] is the first and famous WordNet for English language developed at Princeton University. It delineated the design for the nouns, verbs, adjectives and adverbs of a language to be grouped under sets of synonyms, or synsets. Apart from functioning as a dictionary and thesaurus combined into one, it is used greatly in various NLP applications. English WordNet, in course of time, became one of the most used and valuable language resources. Over a period of time, word nets in other languages got developed along the lines of English WordNet. In case of Indian languages, Hindi WordNet [6] was the first of its kind as far as Indian languages are concerned. Hindi Wordnet was developed at the IIT Bombay.

The Nepali WordNet [8] has been developed at Assam University, Silchar as part of a Consortium Project headed by IIT, Bombay. The Nepali Wordnet is part of the Indo Wordnet [7] Project. The IndoWordNet is a multilingual WordNet which links WordNets of 18 Indian languages viz Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, [1]Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu. The Indo Wordnet Project has linked the synsets of one Indian language to another.

Nepali WordNet is also a machine readable lexical database for the Nepali language along the lines of the famous English WordNet and the Hindi WordNet. The design of the Nepali WordNet is based on the principle of “expansion” from the Hindi Wordnet and English Wordnet. In the Expansion Approach, synsets of a pre-existing WordNet are understood by the lexicographer and the corresponding target language synsets expressing the same sense are created. The synsets have a synset id which is unique. Synset lists the synonymous words in a most frequent order. Synsets are the basic building blocks of WordNet. One sample entry of Nepali WordNet is as follows:

Word: अर्थ

ID: 2971

CAT:NOUN

Synset: अर्थ, अभिप्राय, आशय, मतलब, तात्पर्य, भाव, माने, अंतर्भाव, अन्तर्भाव, अध्यावसान, अरथ, आकूत, आकूती, आसय,

Gloss: वह अभिप्राय या आशय जो किसी शब्द, पद या वाक्य आदि से निकलता है और जिसका बोध कराने के लिए वह शब्द या पद लोक में प्रचलित होता है

Example Sentence: "कभी-कभी सूरदास के पदों का अर्थ निकालना मुश्किल हो जाता है"

Another sense of word अर्थ is

Word: अर्थ

ID: 3051

CAT:NOUN

Synset: धन-दौलत, दौलत, धन, रुपया पैसा, पैसा, वित्त, अर्थ, वैभव, लक्ष्मी, विभव, द्रव्य, इक़बाल, नियामत, ज़र, नेमत, शेव, शुक्र, अरथ, दत्र, अर्बदर्ब, इशरत,

Gloss: रुपया-पैसा, सोना-चाँदी, ज़मीन-जायदाद आदि Example Sentence: "धन-दौलत का उपयोग अच्छे कार्यों में ही करना चाहिए"

3. NEPALI SENSE TAGGED CORPUS

Sense tagging is the task of tagging each word in the sentence with the correct sense of the word [9]. Annotation of words with appropriate senses is very difficult task. To automate the process of sense tagging a Sense Tagger tool has been developed which facilitates the task of manual sense tagging. This tool displays all the senses of the target word available in the Nepali WordNet and allows the annotator to select the correct sense of the word from the candidate senses. Only open-class words nouns, verbs, adjectives, and adverbs are tagged by Sense Tagger tool as Nepali WordNet deals with only open-class words. The corpus from various domains such as media, health, politics etc. is used for sense tagging. A sense annotated corpus is a significant resource for many natural language processing tasks.

The layouts of Sense Tagging Tool are shown in figure 1.

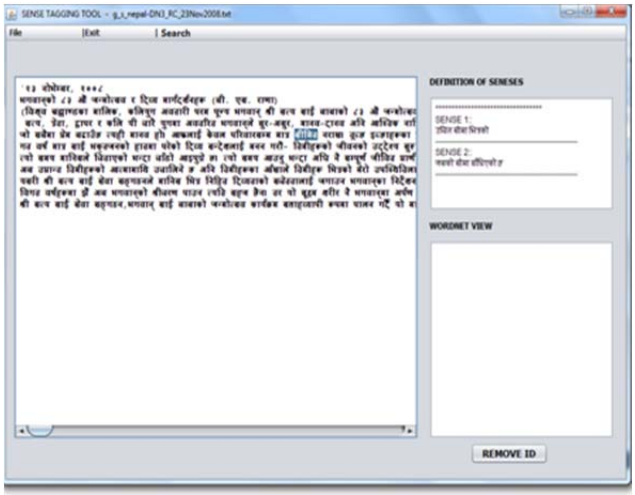

Fig. 1: Layout of Sense Tagger Tool

The sense tagger tool displays all the senses of the word and information such as ID, synset, POS, gloss etc. available in the

wordnet. When the most appropriate sense in the given context is selected from the list of possible senses, synset id corresponding to the synset gets tagged to the word. A screen shot of the sense tagged corpus is shown in the figure 2.

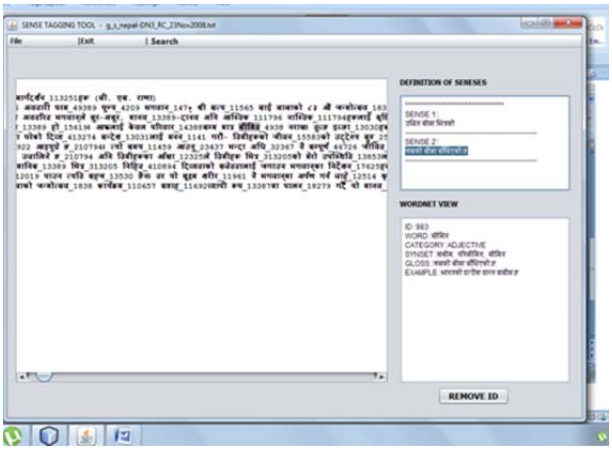

Fig. 2: Sense Tagged Document

Sample sense tagged Nepali text document is shown in table 1

Table 1: Sample sense tagged Nepali documents

आत्महत्या_14698को कारण_118140 स्पष्ट_41450 छ_210794 ? हिमाली राज्य_1231 सिक्किममा प्रायः_31811 दिनहुँ_36459 नै आत्महत्या_14698 भएका घटना_12968हरू प्रकाश_12037मा आउने_415506 गरेका छन्। विशेष_424012 गरी पश्चिम_16616 र दक्षिण_16085 जिल्लामा आत्महत्या_14698का घटना_12968हरू अधिक_42403 हुने_110347 गरेका छन्।

4. CONSTRUCTION OF VECTORS USING SENSE TAGGED DOCUMENTS

In vector space model based on bag of words approach, the component of the document vectors are independent words/terms. The limitations of this approach is that it cannot capture ambiguity, synonymy, semantic relations between words. For example in the sentence, "A bear can bear very cold temperatures", frequency count of word bear is 2. In the above stated example sentence the word "bear" has two different senses, viz., first bear means an animal and meaning of second bear is tolerate. Hence, the problem with VSM using bag of words approach is that it finds the frequency count of a words whereas for a proper generation of document vectors the frequency count of the senses of a word is required. In this work rather than words, sense ids have been used for generation of document vectors. For example consider the sentences, "the fly can fly" and "the stars were shiny and the planets were bright" which appear in a document, the vectors corresponding to these sentences considering words as component of the vectors are shown in the Table 2

Table 2: Document Vector

	Fly	**Star**	**Shiny**	**Planet**	**Bright**
D1	2	0	0	0	0
D2	0	1	1	1	1

Vectors corresponding to the sense tagged sentences of the above example sentences are shown in Table 3. Here synset ids have been considered as the component of the vectors.

Table 3: Document Vector using sense tagged document

	Fly_ 02192818	**Star_ 27080**	**Fly_ 01944262**	**Planet_ 208**	**Shiny_3932& Bright_3932**
D1	1	0	1	0	0
D2	0	1	0	1	2

In the example sentences stated above the word fly has two senses. First *'fly'* is a two- winged insect and second *'fly'* is travel through the air. In the second sentence, the meaning of the words '*shiny'* and '*bright'* are same. The term based method ignores the fact that words may be ambiguous and different words may have same meaning and generate vectors by considering only the frequency count of the words as shown in table 2. We can see from the example that vector generated by the term based method the word *'fly'* has been considered as a single term and has frequency count of 2.Similarly the words *shiny* and *bright* have considered as different words and frequency count of both the words is 1. But in the vector generated from sense tagged sentences fly finds different places in the vector because it has different synset ids and frequency count is 1. Similarly the words *shiny* and *bright* find same place in the vector as they have same id and frequency count is 2. Once the feature vectors are completed in this way, weights are assigned to each word/ synset id across the corpus using TF*IDF method [10], which is the combination of the term/sense frequency (TF), and the inverse document frequency (IDF). Terms which are not present in the WordNet are not considered as element of the feature vector.

5. CLUSTERING ALGORITHMS

In this paper, K-means, Particle Swarm Optimization(PSO), hybrid Particle Swarm Optimization-Kmeans and hybrid Self Organizing Map- Kmeans clustering algorithms are applied to the input vectors for evaluation of vectors generated by sense based method and term based method .

K-means[2] algorithm groups the data vectors into prespecified number of clusters. Initially randomly initialize the centroids of the predefined clusters. The dimension of the centroids are same as the dimension of data vectors. Assignment of each data object to the cluster is done based on the similarity between the data object and the cluster centroid. The reassignment procedure is repeated until the fixed iteration number, or the cluster result does not change after a certain number of iterations.

J. Kennedy and R.C. Eberhart[13] in 1995 first introduced Particle Swarm Optimization model (PSO). PSO consists of a group (swarm) of particles moving in the search space. The particles in the problem space moves with the given velocities according to its own experience and its neighbors' experience In the PSO algorithm, a particle's location in the multi-dimensional problem space represents one solution for the problem. When a particle moves to a new location, a different problem solution is generated. A fitness function is used to evaluate quality of solution.

Hybrid PSO-Kmeans algorithm [14] consists of two modules, the PSO module and the K-means module. At the beginning stage PSO clustering algorithm is executed to find points close to the optimal solution by global search. In the second stage K-means algorithm uses the result of the PSO algorithm as initial centroid vectors. The K-means algorithm is then executed to find the final optimal clustering solution. The hybrid PSO algorithm have the advantage of globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm

Self organization map (SOM)[15] is a type of neural network which is particularly well suited for clustering. The hybrid SOM-Kmeans[16] algorithm is a two level approach for clustering. In the hybrid SOM-Kmeans algorithm first SOM algorithm is applied to input data vectors to produce a set of prototypes or codebook vectors which is much higher than the actual number of clusters and then K-means algorithm is applied to the prototypes vectors to form the actual clusters.

6. EXPERIMENTAL RESULTS

For the evaluation of the vectors generated from sense tagged corpus and word frequency method , four clustering algorithms viz. standard K-Means, Particle Swarm Optimization(PSO), hybrid PSO+K-means and Self Organizing Map(SOM)+K-Means algorithm are used to cluster the vectors. For experiment, Nepali text documents are collected from Technology Development for Indian Language website [11]. The corpus in Nepali language provides data from different domains such as literature, science, media, art etc. The dataset consists of 400 Nepali text documents. The size of the vectors made from sense tagged documents is 2435 and the size of the vectors made by word frequency method is 3174. Cluster quality is assessed by the Silhouette Coefficient.The silhouette coefficient is a measure for the clustering quality that is rather independent from the number of clusters. Silhouette values between 0.7 and 1.0 indicate clustering results with excellent separation between clusters, For the range from 0.5 to 0.7 one finds that data points are clearly assigned to cluster centers. Values from 0.25 to 0.5 indicate that cluster centers can be found, though there is considerable "noise"[12].

For the K-means algorithm no parameter needs to be set up. In the PSO clustering algorithm, 10 random particles are generated. The fitness value is calculated using the distance between the cluster centroid and the documents that are clustered. The inertia weight w was initially set as 0. 95, the acceleration coefficient constants c1 and c2 are set to 1.49 and the iteration number fixed to 50. For the implementation of Self organizing map, SOM Toolbox 2.0, was used.

The silhouette coefficient(SC) values obtained by the application of K-means, Particle Swarm Optimization, Hybrid PSO+K-means and SOM+K-means for word frequency method and sense tagged documents are presented in Table 4.

Table 4: Performance comparison of silhouette coefficient (SC) values

Algorithm	No. Of Clusters	Method	
		Word +VSM	Sense ID +VSM
K-means	3	0.52	0.60
	4	0.27	0.43
	5	0.69	0.47
	6	0.28	0.51
	7	0.40	0.57
	8	0.33	0.41
PSO	3	0.15	0.29
	4	0.22	0.41
	5	0.20	0.36
	6	0.34	0.43
	7	0.22	0.25
	8	0.14	0.18
PSO+K-means	3	0.52	0.67
	4	0.57	0.41
	5	0.25	0.67
	6	0.54	0.71
	7	0.31	0.64
	8	0.43	0.45
SOM+K-means	3	0.24	0.44
	4	0.34	0.60
	5	0.47	0.57
	6	0.40	0.50
	7	0.49	0.69
	8	0.36	0.47

It is observed from the table 4 that the SC value obtained by different clustering algorithms are better for sense tagged documents in almost all the number of clusters compared to the traditional method. However K-Means at number of cluster 5 and hybrid PSO+K-means algorithm at number of cluster 4 in term frequency method of vector generation achieves SC value equal to .69 and .57 which is higher than SC value equal to .47 and .41 respectively achieves using sense tagged document. On an average the experimental result shows that using sense tagged document corpus for generation of vectors improves the clustering quality .

7. CONCLUSION

Sense tagged corpus is very important resource for many research works. In this paper, sense tagged corpus is used for

vector generation. Senses (synset ids) are considered as component of the vectors rather than terms as component of the vector. Four clustering algorithms namely K-means, Particle Swarm Optimization(PSO), hybrid Particle Swarm Optimization-Kmeans and hybrid Self Organizing Map-Kmeans clustering algorithms are applied to the input dataset for evaluation of vectors generated by sense based method and term based method. The experimental results show that clustering quality is better when vectors are generated from sense tagged documents.

REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM

[2] J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, Proceedings of the fifth Berkley Symposium on Mathematical Statistics and Probability, L.M. Lecam J. Neyman (editors), volume 1, Berkley, CA, University of California Press, (1967)281-297.

[3] A. Roy, S. Sarkar, B. S. Purkayastha, "A Proposed Nepali Synset Entry and Extraction Tool," 6th Global wordnet conference, Matsue,Japan,2012

[4] George A. Miller, Richard Beckwith, Christiane Fellbaum,Derek Gross, and Katherine Miller, Introduction to WordNet: An On-line Lexical Database, Princeton University, Cognitive Science laboratory, Technical report, 1993.

[5] C.Felbaum ,"Wordnet : An Electronic Lexical Database" MIT Press,1998

[6] D. Narayan, D. Chakrabarty, P. Pande And P.Bhattacharyya, "An Experience In Building The Indo Wordnet-A Wordnet For Hindi" International Conference On Global Wordnet (GWC 02), Mysore, India, January, 2002.

[7] M. Sinha, M. Reddy, P. Bhattacharyya, "An Approach towards Construction and Application of Multilingual Indo- WordNet", 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006

[8] A. Chakraborty, B. S. Purkayastha, A. Roy, "Experiences in building the Nepali Wordnet" Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India, Mumbai 2010

[9] A. Chatterjee, S. R. Joshi, M. M. Khapra and P. Bhattacharyya 2010. Introduction to Tools for IndoWordnet and Word Sense Disambiguation, The 3rd IndoWordnet Workshop, Eighth International Conference on Natural Language Processing (ICON 2010), IIT Kharagpur, India

[10] J. Sedding and D. Kazakov, "WordNet-based Text Document Clustering," ROMAND, page104, 2004.

[11] http://tdil-dc.in.

[12] T Gharib, M. M Fouad, A. Mashat, I.Bidawi. Self Organizing Map based Document Clustering Using WordNet Ontologies. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 2, 2012

[13] J. Kennedy and R.C. Eberhart, "Particle Swarm Optimization," Proc. IEEE, International Conference on Neural Networks. Piscataway. Vol. 4, pp 1942-1948,1995

[14] Cui, T.E. Potok, Document Clustering Analysis Based on Hybrid PSO+ K-Means Algorithm, Journal of Computer Sciences (Special Issue): 27-33, 2005 ISSN 1549-3636

[15] T. Kohonen, : Self-organized formation of topologically correct feature maps. Biol.Cybern. 43 59–69 (1982)

[16] J. Vesanto and E. Alhoniemi, Clustering of the Self-Organizing Map, IEEE Transactions On Neural Networks, Vol. 11, No. 3, May 2000