A Comparative Analysis of Particle Swarm Optimization and K-means Algorithm For Text Clustering Using Nepali Wordnet

Sunita Sarkar[1] , Arindam Roy[2] and B. S. Purkayastha[3]

Department of Computer Science, Assam University, Silchar, Assam, India

ABSTRACT

The volume of digitized text documents on the web have been increasing rapidly. As there is huge collection of data on the web there is a need for grouping(clustering) the documents into clusters for speedy information retrieval. Clustering of documents is collection of documents into groups such that the documents within each group are similar to each other and not to documents of other groups. Quality of clustering result depends greatly on the representation of text and the clustering algorithm. This paper presents a comparative analysis of three algorithms namely K-means, Particle swarm Optimization (PSO) and hybrid PSO+K-means algorithm for clustering of text documents using WordNet. The common way of representing a text document is bag of terms. The bag of terms representation is often unsatisfactory as it does not exploit the semantics. In this paper, texts are represented in terms of synsets corresponding to a word. Bag of terms data representation of text is thus enriched with synonyms from WordNet. K-means, Particle Swarm Optimization (PSO) and hybrid PSO+K-means algorithms are applied for clustering of text in Nepali language. Experimental evaluation is performed by using intra cluster similarity and inter cluster similarity.

.

KEYWORDS

Nepali WordNet; Text clustering; Particle Swarm Optimization; K-means.

1. INTRODUCTION

Digitized text documents is increasing exponentially. As such, clustering becomes imperative for ever increasing digitized data. Clustering will facilitate applications like document organization, data analysis, information retrieval and so on. Clustering is an useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups [22]. Clustering is one of the important techniques of data mining. Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data[1]. Text (document) Clustering is defined as the collection into groups such that the documents within each group are similar to each other and dissimilar to those in other groups. Two important factors on which quality of clustering result depends are

1. Representation of text document
2. Choice of clustering algorithm.

Representation of texts is vital for tasks like retrieval, classification, clustering, summarization, question-answering and so on[2]. Vector Space Model with bag of words representation is the common and simple way of representing a text. This representation has certain drawbacks like it does not exploit semantic relations between the words. The “bag of words” representation is a

10.5121/ijnlc.2014.3308

semantic neutral approach which treats close semantic relation and no semantic relation between two terms uniformly with the consequence that quality of clustering result can get significantly degraded. In this paper an effort has been made to improve the clustering by representing a text using semantics relation. Distinct terms share the same meanings are known as synonym. Set of synonyms stand for concepts which correspond to the words of the text[2]. In this work feature vector is generated using WordNet. K-means, Particle swarm Optimization (PSO) and hybrid PSO+K-means clustering algorithms are applied to those feature vectors to cluster the text documents in Nepali[1] language. Several experiments have been performed to analyze the effects of these clustering methods on Nepali text document dataset using cosine similarity measurement.

The rest of this paper is organized as follows: Section II provides a description of Nepali WordNet. Related work is discussed in section III. Section IV is devoted to methodology. In section V clustering algorithms are discussed. Section VI discusses the experimental results and Section VII concludes the paper.

.

2. NEPALI WORDNET

The Nepali WordNet [4] has been developed at Assam University, Silchar as part of a Consortium Project headed by IIT, Bombay with a generous grant from Technology Development of Indian Language Programme, Department Of Information Technology, Ministry of Communications and Information Technology, India. It is a machine readable lexical database for the Nepali language along the lines of the famous English Wordnet and the Hindi Wordnet. Nepali WordNet is a system for bringing together different lexical and semantic relations between the Nepali words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Nepali WordNet is based on the principle of "expansion" from the Hindi WordNet and English WordNet.

Four major components of WordNet are ID,CAT, SYNSET and GLOSS. ID is the unique synset identifier. CAT specifies the following category of words- Noun, Verb, Adjective and Adverb. SYNSET lists the synonymous words in a most frequent order. Synsets are the basic building blocks of WordNet and GLOSS describes the concept of any synset[5].It consists of Text-Definition and Example- Sentence. Text-Definition contains concept denoted by synset and Example illustrates the use of any word in synset list. One sample entry of Nepali WordNet is as follows:

ID:8231

CAT:NOUN

Synset: **भण्डार**, **भँडार**, **ढुकुटी**, **सागर**, **समुद्र**

Gloss: कुनै विषयको ज्ञान वा गुण आदिको धेरै ठुलो ढुकुटी

Example Sentence:"सन्त कबीर ज्ञानका भण्डार थिए"

Various semantic relations that exist between the synsets of WordNet are: Hyponym/Hypernym, Meronym/Holonym, Entailment/Troponym.

3. RELATED WORK

Text document clustering can be challenging due to complex linguistics properties of the text documents. Most of the clustering techniques are based on traditional bag of words approach to represent the documents. In such document representation, ambiguity, synonymy and semantic similarities may not be captured using traditional text mining techniques that are based on word and/or phrase frequencies in the text. Gad and Mohamed S. Kamel [6] proposed a semantic similarity based model to capture the semantic of the text. The proposed model in conjunction with lexical ontology solves the synonyms and hypernyms problems. It utilizes WordNet as an ontology and uses the adapted Lesk algorithm to examine and extract the relationships between terms. The proposed model reflects the relationships by the semantic weighs added to the term frequency weight to represent the semantic similarity between terms.

In [7,8] authors introduced conceptual features in text representation. A concept feature is an aggregation of a few words that describe the same high level concept, for example, dog and cat describing the concept animal. They proposed three methods to include concept features in VSM, namely, (i) adding concept features to the term space (i.e., term+concept); (ii) replacing the related terms with concept features and (iii) reducing the VSM to only concept features. For text clustering, experimental results [8] showed that only the term+concept representation improved clustering performance.

Sridevi and Nagaveni [9] showed that combination of ontology and optimization to improve the clustering performance. They proposed a ontology similarity measure to identify the importance of the concepts in the document. Ontology similarity measures is defined using wordnet synsets and the particle swarm optimization is used to cluster the document.

In [10] authors proposed a semantic text document clustering approach based on the WordNet lexical categories and Self Organizing Map (SOM) neural network. The proposed approach generates documents vectors using the lexical category mapping of WordNet after preprocessing the input documents.

Liping Jing et al[11] proposed a new similarity measure combining the edge-counting technique and the position weighting method to compute the similarity of two terms from an ontology hierarchy. They modified the VSM model by readjusting term weights in the document vectors based on its relationships with other terms co-occurring in the document. They applied three different clustering algorithms, bisecting k-means, feature weighting k-means and a hierarchical clustering algorithm to cluster text data represented in knowledge based VSM.

[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].

4. METHODOLOGY

To apply any clustering algorithms on text document dataset, it is necessary to generate document vector from document dataset. So proper representation of document is very crucial so that they can be represented easily and reduces complexity. The figure1 below describes the approach.

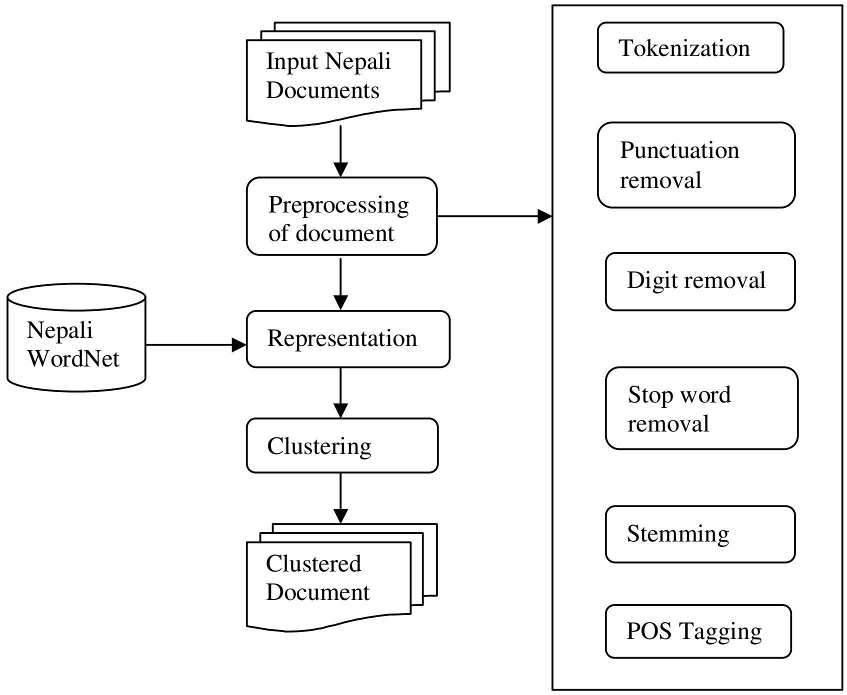

Fig1: Steps Showing Clustering of Text

4.1. Preprocessing of Text Document

Preprocessing of documents is a vital step for the quality of clustering result. In this work we are dealing with the Nepali documents. Obviously these texts will have sufficient number of Nepali punctuations, Nepali and English digits and Single-letter words. These are very common and contribute nothing to the actual content of the text. While applying clustering algorithm to document dataset we need not take into consideration all these unnecessary characters. So during this phase we remove these characters step by step. Preprocessing consists of the following steps:- i)Tokenization, ii)Punctuation removal iii)Digit removal iv)Stop word removal v)Stemming and vi)Part of speech tagging An example with the following input sentence demonstrates this.

भारत कम्युनिस्ट पार्टी (मार्क्सवादी) को १९ – औँ कङ्ग्रेस कोयम्बटुरमा समाप्त भयो।

Tokenization : Tokenization is the process of breaking the sentences into individual tokens. The words within [ ] are the tokens. For instance
[भारत] [कम्युनिस्ट] [पार्टी] [(] [मार्क्सवादी] [)] [को] [१९] [–] [औँ] [कङ्ग्रेस] [कोयम्बटुरमा] [समाप्त] [भयो] [।]

Punctuation Removal: As we are working on general Nepali text, there are lots of Nepali punctuations in the text. These characters have no importance. So we removed these punctuations. For instance

भारत कम्युनिस्ट पार्टी मार्क्सवादी को १९ औँकङ्ग्रेस कोयम्बटुरमा समाप्त भयो

Digit Removal: A general Nepali text file may contain Nepali as well as English digits. But as meaningful Nepali words do not contain digits, we remove these digits.
भारत कम्युनिस्ट पार्टी मार्क्सवादी को औँ कङ्ग्रेस कोयम्बटुरमा समाप्त भयो

Stop Word Removal: There exist a lot of words having a single letter. Most of these Single-Letter-Words are Stop- Words. We removed stop words in this phase.
भारत कम्युनिस्ट पार्टी मार्क्सवादी कङ्ग्रेस कोयम्बटुरमा समाप्त भयो

Stemming: After removing all stop words we stemmed them to a common lexical root using Nepali Stemmer[12].
भारत कम्युनिस्ट पार्टी मार्क्सवादी कङ्ग्रेस कोयम्बटुर समाप्त भयो

Parts of speech tagging: PoS tags are assigned to the text document using PoS tagger[13]. For instance
भारत/NP कम्युनिस्ट/NN पार्टी/NN मार्क्सवादी/NN कङ्ग्रेस/NN कोयम्बटुर/NN समाप्त/JX भयो/VVYN

4.2. Document Representation Based on WordNets

In the vector space model a document d is represented as n-dimensional document vector $[wt_0, wt_1, \ldots, wt_n]$, where t_0, t_1, . . .,t_n is a set of distinct terms present in given document and wt_i expresses the weight of term t_i in document d[15]. The weight of a term reflects the importance of term within a particular document. Nepali WordNet has been used to represent Nepali documents. Nepali WordNet is a lexical database that provides the sense information. A term may have more than one sense. In WordNet, ID is unique for each synset for each sense. All the synonyms in this synset share the same ID. Sense tagged data have been used to extract the exact sense of a word .For the synset concept, corresponding ID of the word is taken and vectors of IDs are prepared. Once the document vectors are completed in this way, weights are assigned to each word across the corpus using TF*IDF method[14], which is the combination of the term frequency (TF), and the inverse document frequency (IDF). TF*IDF is mathematically written as

$$W_{ij} = tf_{i,j} * \log (N / df_i)$$

Where W_{ij} is the weight of the term **i** in document **j**.
$tf_{i,j}$ is the number of occurrences of term **i** in document **j**.
N is the total number of documents in the corpus,
df_i is the number of documents containing the term **i**.

4.3. Similarity Measure

Similarity between two documents need to be computed in a clustering analysis. There are several similarity measures are available in the literature to compute the similarity between two documents like Euclidean distance, Manhattan distance, cosine similarity etc. Among these measurements, cosine similarity measure[20] has been used to compute the similarity between two documents in the experiments.

$$cos(d_1,d_2) = \frac{(d_1 \bullet d_2)}{\| d_1 \| \| d_2 \|} \qquad (1)$$

where · and || indicates dot product and length of a vector respectively.

5. Clustering Algorithm

5.1. K-Means Algorithm

In K-means algorithm data vectors are grouped into predefined number of clusters. At the beginning the centroids of the predefined clusters are initialized randomly. The dimension of the centroids are same as the dimension of data vectors. Each data object is assigned to the cluster 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.

The K-means algorithm is summarized as

1. Randomly initialize the N_c cluster centroid vectors
2. Repeat
   (a) For each data vector, assign the vector to the class with
      the closest centroid vector,
   (b) Recalculate the cluster centroid vectors, using

$$c_j = \frac{1}{n_j} \sum_{\forall d_j \in S_j} d_j \qquad (2)$$

where d_j denotes the document vectors that belong to cluster S_j; c_j stands for the centroid vector; n_j is the number of document vectors that belong to cluster S_j.
until a stopping criterion is satisfied.

5.2. Particle Swarm Optimization

PSO is an evolutionary computation technique first introduced by Kennedy and Eberhart in 1995 [16]. PSO is a population-based stochastic search algorithm which is modeled after the social behavior of a bird flock. In the context of PSO, a swarm refers to a number of potential solutions to the optimization problem, where each potential solution is referred to as a particle. The aim of the PSO is to find the particle position that results in the best evaluation of a given fitness (objective) function [17].

Each individual in the particle swarm is composed of three D-dimensional vectors, where D is the dimensionality of the search space. These are the current position x_i, the previous best position p_i, and the velocity v_i [18].The i^{th} particle is represented by a position denoted as $x_i = (x_{i1}, x_{i2},. . . , x_{iD})$. In a PSO system, each particle flows through the multidimensional search space, adjusting its position in search space according to its own experience and that of neighboring particles. To evolve towards an optimal solution a particle uses a combination of the best position realized by itself and the best position realized by its neighbours. The standard PSO method updates the velocity and position of each particle according to the equations given below.

$$v_{id}(t+1)=\omega.v_{id}(t)+c_1.rand().(p_{id}-x_{id})+c_2.rand().(p_{gd}-x_{gd}) \quad (3)$$
$$x_{id}(t+1)=v_{id}(t+1)+x_{id}(t) \quad (4)$$

where c_1 and c_2 are two positive acceleration constants, rand() is a uniform random number in (0, 1), p_{id} and p_{gd} are the best positions found so far by the ith particle and all the particles respectively, t is the iteration count and ω is an inertia weight which is usually, linearly decreasing during the iterations. The inertia weight ω plays a role of balancing the local and global search.

In the context of clustering, a single particle represents the N_c cluster centroid vectors. That is, each particle x_i is constructed as follows:

$$x_i=(o_{i1},\ldots,o_{ij},\ldots,o_{iNc}) \quad (5)$$

Where o_{ij} refers to the jth cluster centroid vector of the ith particle in cluster C_{ij}. Therefore, a swarm represents a number of candidate clusters for the current data vectors. The fitness of particles is measured using the equation given below.

$$f=\frac{\sum_{i=1}^{N_c}\{\frac{\sum_{j=1}^{P_i} d(o_i,m_{ij})}{P_i}\}}{N_c} \quad (6)$$

where m_{ij} denotes the jth document vector, which belongs to cluster i; o_i is the centroid vector of the ith cluster; $d(o_i, m_{ij})$ is the distance between document m_{ij} and the cluster centroid o_i; P_i stands for the number of documents, which belongs to cluster C_i and N_c stands for the number of clusters.

The PSO Clustering algorithm can be summarized as:

```
(1) Initially, each particle randomly select k different document vectors from the document
collection as the initial cluster centroid vectors.
(2) For t= 1 to $t_{max}$ do
   a)For each particle i do:
   b)For each document vector  $m_p$ do
     (i) Calculate the distance d ($m_p$,$o_{ij}$), to all cluster centroids $C_{ij}$
     (ii) Assign each document vector to the closest centroid  vector.
     (iii) Calculate the fitness value based on equation (6).
   c) Update the global best and local best positions
d) Update the cluster centroids using equations (3) and (4)
```

Where t_{max} is the maximum number of iterations.

5.3. Hybrid PSO + K-Means Algorithm

In the hybrid PSO algorithm [19], the algorithm includes two modules, the PSO module and the K-means module. The hybrid algorithm first executes PSO clustering algorithm to find points close to the optimal solution by global search and simultaneously avoid high computation time. In this case PSO clustering is terminated when the maximum number of iterations is exceeded. The result of the PSO algorithm is then used as initial centroid vectors of the K-means algorithm. The K-means algorithm is then executed until maximum number of iterations is reached. The K-means algorithm tends to converge faster (after less function evaluations) than the PSO, but usually with a less accurate clustering [23] and PSO can conduct a globalized searching for the optimal clustering, but requires more iteration numbers and computation than the K-means algorithm. The hybrid PSO algorithm combines the advantage of both the algorithms: globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm**.**

6. EXPERIMENTAL RESULT

In this paper, the experiments were conducted on Nepali text document datasets with the K-means, PSO and hybrid PSO+K-means algorithms. Nepali text dataset has been collected from Technology Development for Indian Language website [21]. The corpus in Nepali language provides data from different domains such as literature, science, media, art etc. Sense tagging of the whole document corpus was done using sense marking tool. Document preprocessing step is implemented in Java using NetBeans 7.3 and clustering algorithm (hybrid PSO+K-means) is implemented in MATLAB (Version 7.9.0.324). All experiments were done on Processor P4 (3GHz) machine with 2GB main memory, running the Windows 7 Professional® operating system. In the PSO clustering algorithm, we choose 10 particles, the inertia weight ***w*** is initially set as 0.95 and the acceleration coefficient constants ***c1*** and ***c2*** are set as1.49. In this study the quality of the clustering is measured according to the following two criteria:

Intra-cluster similarities, i.e. the distance between data vectors within a cluster, where the objective is to maximize the intra-cluster similarity;

Inter-cluster similarity, i.e. the distance between the centroids of the clusters, where the objective is to minimize the similarity between clusters.

These two objectives respectively correspond to crisp, compact clusters that are well separated. Both intra-cluster similarity and inter-cluster similarity are internal quality measures. The results obtained are shown in table 1.

Table 1. Performance comparison of K-Means, PSO and Hybrid PSO+ K-Means Algorithms

No. of document	Method	Algorithm	Intra cluster	Inter cluster
50	Term Based	K-means	7.0931	.5855
		PSO	6.9646	.9522
		Hybrid PSO+K-means	7.1258	.5204
	Sense Based	K-means	7.1043	.5682
		PSO	7.0419	1.053
		Hybrid PSO+K-means	7.9640	.5171

7. CONCLUSION

A comparative analysis of three algorithms namely K-means, PSO and hybrid PSO+K-means has been presented for document clustering problems. The K-means algorithm was compared with PSO and hybrid PSO+K-means algorithms. In this work we used a lexical database in the form of Nepali WordNet to represent the text documents with sense/ concept. Synsets in Nepali WordNet provided semantically related terms. Sense tagged corpus is used to extract correct sense of a particular term. Experimental results demonstrate that and hybrid PSO+K-means performs better than PSO and K-means algorithms when documents are represented using WordNet.

REFERENCES

[1] R. Bhagel, and R. Dhir "A Frequent Concept Based Document Clustering Algorithm", IJCA, vol 4, no.5, 2010.

[2] G. Ramakrishnan and P. Bhattacharyya, "Text Representation with Wordnet Synsets", Eight International Conference on Applications of Natural Language to Information Systems (**NLDB2003**), Burg, Germany, June, 2003.

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

[4] 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

[5] J.Sarmah, A.K. Barman, and S.K. Sarma, "Automatic Assamese Text Categorization Using WordNet", International conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, 978-1-4799-2432-5 IEEE, pp 85-89,Mysore, India 2013

[6] W. K. Gad and M. S. Kamel, "Enhancing Text Clustering Performance Using Semantic Similarity", ICEIS, LNBIP 24, pp. 325–335, 2009

[7] A. Hotho, S. Staab, G. Stumme, "Wordnet improves text document clustering". In: Proceedings of the semantic web workshop at 26th annual international ACM SIGIR conference, Toronto, Canada, 2003

[8] A. Hotho, S. Bloehdorn "Text classification by boosting weak learners based on terms and concepts".In Proceedings of the IEEE international conference on data mining, Brighton, UK, pp 72-79,2004.

[9] U. K. Sridevi and . N. Nagaveni. "Semantically Enhanced Document Clustering Based on PSO Algorithm", European Journal of Scientific Research, ISSN 1450-216X Vol.57 No.3, pp.485-493,2011

[10] T.F Gharib, M. M Fouad, A. Mashat, I.Bidawi1 " Self Organizing Map based Document Clustering Using WordNet Ontologies". IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 2, 2012

[11] L. Jing, K. Ng. Michael, J. Z. Huang, " Knowledge-based vector space model for text clustering ", Knowl Inf Syst 25:35–55, 2010

[12] B. Krishna , P.l Shrestha,"A Morphological Analyzer and a stemmer for Nepali,"
www.nepalinux.org/downloads/nlp/stemmer_ma.zip.

[13] M.R.Jaishi, "Hidden Markov Model Based Probabilistics Part of Speech Tagging For Nepali Text, Masters Dissertation. 2009

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

[15] D. Weiss, "Descriptive Clustering as a Method for Exploring Text Collections," Ph.D Thesis.

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

[17] S.C. Satapathy, N. VSSV P B. Rao, JVR. Murthy, R. P.V.G.D. Prasad, "A Comparative Analysis of Unsupervised K-means, PSO and Self- Organizing PSO for Image Clustering," International Conference on Computational Intelligence and Multimedia Applications,2007.

[18] S. Sarkar, A.Roy, B.S.Purkayastha, “Application of Particle Swarm Optimization in data clustering : A survey,” International Journal Of Computer Applications (0975- 8887) Volume 65- No.25, 2013

[19] X. Cui, T.E. Potok, P. Palathingal, “Document Clustering using Particle Swarm Optimization,” IEEE,2005

[20] X. Rui, "Survey of Clustering Algorithms", IEEE transactions on Neural Networks, 16(3), pp.634-678, 2005

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

[22] A. Huang,"Similarity Measures for Text Document Clustering" *NZCSRSC 2008*, April 2008, Christchurch, New Zealand.

[23] M. Omran, A. Salman, A.P. Engelbrecht, “Image Classification using Particle Swarm Optimization,” Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning, Singapore, 2002