ONLINE CLUSTERING USING K-MEANS and SOM

K.Chandru[#1], K.Duraiswamy[*2], B.Mahalakshmi[#3]

CSE-CSE Department, KSR College of Technology- KSR College of Technology
Tiruchengode, India
[1]chandruraja01@gmail.com
[3]maharajan2203@gmail.com

[*]*Dean of CSE Dept*
KSR College of Technology
Tiruchengode, India

Abstract: **Document clustering is the process of document organization, extraction and information retrieval. Document clustering can be done both in online and offline. Online document clustering provides many irrelevant results, so k-means and self organizing map (SOM) can be used. In k-means algorithm, initially k initial documents as clusters and assign the centroids of these clusters. And text data is directional, and assign document vectors to each cluster. The resulting algorithm is called online spherical k-means. On the other hand SOM can be used to organize input documents into 2-D representation. The resultant document can compared with k-means and SOM on the basis of performance of clustered data. Some of datasets can used to compare the given input documents and predefined datasets for efficient information retrieval.**

Keywords: **k-means, SOM, pre-processing, cloud datasets.**

I.INTRODUCTION

Online users to find the useful information from these huge amounts of data through internet. Search engines and recommender systems help people to reduce the information overload by finding relevant information on their search topic. Clustering of documents is one of the techniques used in search engines and in recommender systems for efficiently finding documents that have similar topics, for improving the performance of information retrieval systems, for assisting users on a web site and for personalization of search engine results. Formally, document clustering is an optimization problem where the input of the problem is a set of documents and a similarity measure between these documents. Thus, similarity plays an important role in document clustering.

Clustering provides an effective navigation mechanism to organize this large amount of data by grouping their documents into a small number of meaningful classes. Text document clustering can be defined as the process of grouping of text documents into semantically related groups. Most of the current methods for text clustering are based on the similarity between the text sources. Various algorithms can be used to have efficient text clustering.

II.STRUCTURED REPRESENTATION OF DOCUMENT FEATURES

In this section describe in details the components of the proposed k-means text clustering approach. There are two main processes: Document Pre-processing that generated output document vectors from input text documents using several processes. The second step is Document Clustering that applies SOM neural network on the generated document vectors to obtain output clusters. For clustering, datasets can be accessed through cloud; so that huge information can retrieve [1].No need have a database to extract predefined datasets. The main advantage for using datasets in cloud is time taken to retrieve a document is reduced.

Document Pre-processing:

Document pre-processing which aims to represent the corpus (input documents collection) into vector space model. Data pre-processing is a very important and essential phase in an effective document clustering. The first part of feature extraction is pre-processing the lexicon and involves removal of stop words and stemming [2]. The stop words removal accounts to 20% to 30% of total words counts while the process of stemming reduces the number of terms in the document. Both the process helps in improving the effectiveness and efficiency of text processing as they reduce the indexing file size.

Filtering:

The process of removing special characters and punctuation that are not thought to hold any discriminative power in the given document. In the formatted documents, such as web pages of different websites, where formatting tags can either be discarded or identified and each has different weights.

***Stemming**:*

The process of a reducing words to their base. For example, the words connects, connected, connection are all reduced to the stem connect.

III. Document Clustering

Clustering is one technology to find intrinsic structures in data sets. Text clustering method usually uses the document vector space model to split the document into vectors in high dimensional space, and then make clustering of these vectors. Text clustering can generally be divided into partitioned clustering algorithms and hierarchical clustering algorithms.

Provide separate documents vectors for all the input documents using feature extraction process.

The problem of document clustering is defined as follows. Given a set of n documents called DS1, DS1 is clustered into a user-defined number of k document clusters D1,D2,…Dk, (i.e. {D1, D2,…Dk} = DS1) so that the documents in a document cluster are similar to one another while documents from different clusters are dissimilar. In this stage apply three different clustering algorithms which are k-means (partitioning clustering) and SOM neural network. These algorithms are most commonly used in the document clustering [4].

A. K-means:

The k-means algorithm start by choosing k initial documents as clusters, and iteratively assign documents to clusters while updating the centroids of these clusters. Advantages of this method are that a number of clusters need not be supplied in advance, and that the resulting cluster is browsable. The resulting algorithm is called spherical k-means [5].

Self-Organizing Map

A basic SOM consists of neurons located on a regular low-dimensional grid that is usually in 2-D. The lattice of a 2-Dgrid is either hexagonal or rectangular. Each neuron has a dimensional feature vector. During the training process, the neurons are updated in a way that their feature vectors finally become representative of different data clusters in an orderly fashion. Self-organizing maps (SOM) learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighbouring neurons in the self-organizing map learn to recognize neighbouring sections of the input space. It Focus on using SOM to perform the document clustering. The two reasons for using SOM rather than other clustering methods are that it is topologically preserving and clustering is performed nonlinearly on the given input data sets. The topologically preserving property allows the SOM applied to document clustering, to group similar documents together in a cluster and organize similar clusters close together unlike most other clustering methods [3].In this approach, use the implementation of Self organizing maps in MATLAB (*Neural Network Toolbox*). Construct a 1-D SOM neural network that takes the generated document vector as input. The size of the network (number of hidden neurons) is based on the desired number of clusters. The network then is trained on the input document vector for about 250 epochs. The output from the network is the weights that define the centers of each cluster. Then assign each document into its appropriate cluster to be evaluated after that.

IV.CONCLUSION

The text document clustering approach based on the online spherical k-means and SOM. The proposed approach can generates high accurate documents by comparing k-means and SOM on the basis of performance of clustered data. The output clusters in each case are compared with datasets which has been stored in cloud. The performance of the proposed approach can be increased by using these datasets. The k-means can achieves higher clustering quality when it extends to online spherical k-means. Also, by including these Types of datasets provide the feature extraction process for text documents improves the overall clustering quality. Finally, the proposed approach shows good scalability against the huge number of documents as in datasets in cloud along with different values of desired clusters.

ACKNOWLEDGMENT

Chandru K is a Master of Engineering Student in the Computer Science and Engineering Department, K.S.Rangaswamy College of Technology. He research interests are Soft computing (neural networks).The basic comparison of k-means and SOM can be done.

REFERENCES:

[1] S. Deng and H. Peng, “Document classification based on support vector
machine using a concept vector model,” in *Proc. IEEE/WIC/ACM Int.Conf. Web Intell.*, Dec. 2006, pp. 473–476.
[2] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining Knowl. Disc.*, vol. 2, no. 2, pp. 121–167,1998.
[3] T. Onoda, H. Murata, and S. Yamada, “SVM-based interactive document
retrieval with active learning,” *New Gen. Comput.*, vol. 26, no. 1,pp. 49–61, 2008.
[4] R. Nallapati, “Discriminative models for information retrieval,” in*Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*,2004, pp. 64–71.
[5] R. T. Freeman and H. Yin, “Web content management by self-organization,”
IEEE Trans. Neural Netw., vol. 16, no. 5, pp. 1256–1268, Sep.2005.
[6] K. Monostori, A. Zaslavsky, and H. Schmidt, “MatchDetectReveal:
Finding overlapping and similar digital documents,” in *Proc. Inf. Resources
Manage. Assoc. Int. Conf. Challenges Inf. Technol. Manage.21st Century*, Anchorage, AK, 2000, pp. 955–957.
[7] S. B. J. Davis and H. Garcia-Molina, “Copy detection mechanisms fordigital documents,” in *Proc. ACM SIGMOD Annu. Conf.*, 1995, pp.398–409.
[8] N. Shivakumar and H. Garcia-Molina, “Building a scalable and accuratecopy detection mechanism,” in *Proc. 1st ACM Conf. Digit. Libraries*,Bethesda, MD, 1996, pp. 160–168.
[9] A. Si, H. V. Leong, and R. W. H. Lau, “CHECK: A document plagiarism
detection system,” in *Proc. ACMSymp. Appl. Comput.*, Feb. 1997,pp. 70–77.
[10] R. B. Yates and B. R. Neto, *Modern Information Retrieval*. Reading,
MA: Addison-Wesley/Longman, 1999.