Volume 5, Issue 2, February 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Bisecting K-means Algorithm for Text Clustering

Nikita P. Katariya, Prof. M. S. Chaudhari
Dept. of Computer Science & Engg
PBCE, Nagpur, India

***Abstract*— Text mining is research technologies to discover useful knowledge from enormous collections of documents and to develop a system to provide knowledge and to support in decision making. Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. Text clustering has become a problem in recent years because of the incredible amount of unstructured data is available in various forms on the web, social networks, and other information networks. This paper gives idea about the bisecting k-means algorithm used for clustering text data.**

Keywords—text mining, text clustering, bisecting k-means algorithm

I. INTRODUCTION

Text mining refers to the process of deriving high-class information from text. High-class information is typically derived through the plan of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text deriving patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, document summarization, and entity relation modelling. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted. Text clustering plays an important role in text mining process. For clustering different methods are used such as partition based clustering, density based clustering or hierarchical clustering method. In this paper we give clustering using bisecting k-means algorithm which is a combination of k-Means and hierarchical clustering.

II. PROCESS OF TEXT MINING

Text Mining is the process of extracting interesting information, knowledge or patterns from the unstructured text that are from different sources. As the text is in unstructured form, it is quite difficult to deal with it. Finding patterns of interested information from the natural language text is the purpose of text mining. The Text Mining Process is shown in Fig. 1

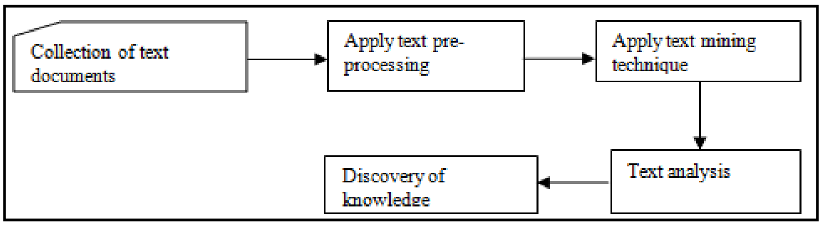
Collection of text documents
Apply text pre-processing
Apply text mining technique
Text analysis
Discovery of knowledge

Fig1. Text mining process

Step1- Pre-processing Text: Mining from a pre-processed text is easy as compare to natural languages documents. So, pre-processing of documents that are from different sources is an important task during text mining process before applying any text mining technique.
Step2- Application of Text Mining Technique: This is an important stage in which the selected algorithm is applied on text in order to process the text. The algorithm such as clustering, classification, summarization, information extractions can be used.
Step3 - Analysis of Text: In this step outputs are analysed for discovering the knowledge. Various tools such as link discovery tool can be used or the outputs can be visualised so that the users can use them for their purpose

III. CLUSTERING

Clustering is a division of data into groups of similar objects. Each cluster consists of objects that are similar to each other and dissimilar to objects of other clusters. The goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances. Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labelled by hand, and then replicates the learnt behaviour on unlabeled data. The applications of clustering are Finding Similar Documents, organizing Large Document Collections, duplicate Content Detection, recommendation System and Optimization of Search.

IV. BISECTING K-MEANS ALGORITHM

This method is a type of hierarchical clustering method using k-means. The algorithm starts by putting all the documents in a single cluster. It partitions the original cluster into two clusters by using K-Means i.e. K = 2. It makes the cluster which has highest intra cluster similarity as permanent and recursively split the other cluster into two more clusters using K-means with K=2 and continue this until the desired number of clusters are created.

A. Algorithm Bisecting K-Means

```
Input: K: Number of clusters, D: Top N documents obtained by vector space similarity
Output: K clusters
put all the N documents in a single cluster C
        for i=1 to K-1 do
          for j=1 to ITER do
                  Use K-means to split C into two sub-clusters, C1 and C2
                   if ( intra-cluster similarity(C1) > intra-cluster similarity(C2) )
                         make cluster C1 as permanent
                          C = C2
                  else
                         make cluster C2 as permanent
                        C = C1
                end if
         end for
     end for
end Bisecting K-Means
```

The critical part is which cluster to choose for splitting. And there are different ways to proceed, for example, you can choose the biggest cluster or the cluster with the worst quality or a combination of both. Fig 2. shows working of bisecting k-means algorithm.

B. Time Complexity

Bisecting K-Means uses K-Means to compute two clusters with K=2. As K-Means is O(N), the run time complexity of the algorithm will be O((K-1)IN), where I is the number of iterations to converge. Hence Bisecting K-Means is linear in the size of the documents.

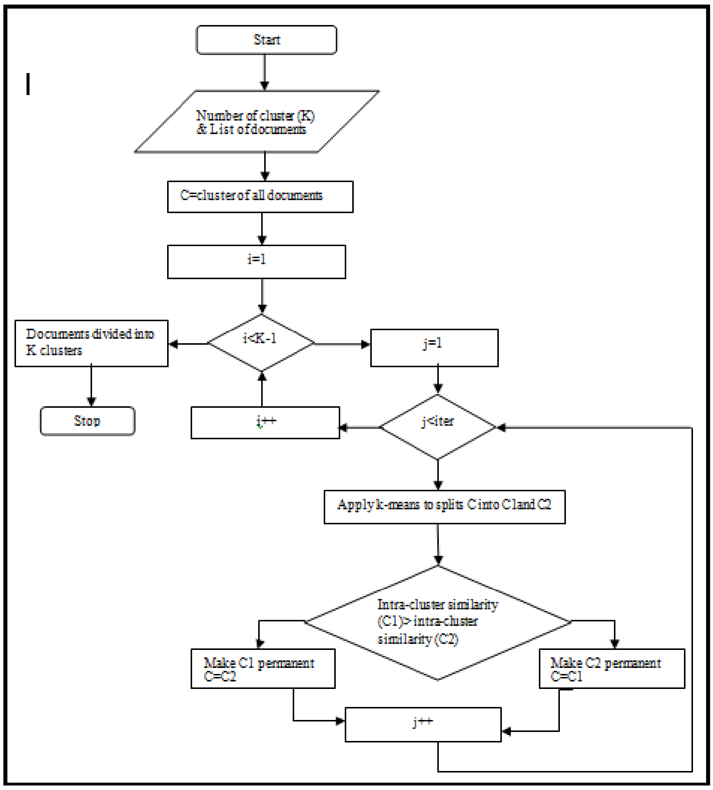

Fig2. Working of bisecting k-means algorithm

V. CONCLUSION

Text data clustering arises in the context of many application domains. Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. This paper presents idea about clustering text documents using bisecting k-means algorithm. Our study indicates that the bisecting K-means technique is better than the standard K-means approach and it produces better clustering solutions according to the dissimilarity and overall similarity measures of cluster quality.

REFERENCES

[1] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.

[2] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

[3] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012, pp. 30-44.

[4] C. C. Aggarwal and C.-X. Zhai, “A survey of text classification algorithms,” in Mining Text Data. New York, NY, USA: Springer, 2012.

[5] R. Sagayam, S.Srinivasan, S. Roshni, “A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques”, International Journal of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, September 2012,pp.1443-1446

[6] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, “An evaluation of feature selection for text clustering”, in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488–495.

[7] Zdenek Ceska and Chris Fox, “The Influence of Text Pre-processing on Plagiarism Detection”, International Conference RANLP 2009 - Borovets, Bulgaria, pages 55–59

[8] Michael Steinbach, George Karypis, and Vipin Kumar, “A Comparison of Document Clustering Techniques”, Department of Computer Science and Egineering,University of Minnesota, Technical Report #00-034.

[9] R.Indhumathi, and Dr.S.Sathiyabama, “ Efficient time reduction using principal component analysis with bisecting k means algorithm”, international journal of engineering science and Technology, Vol. 5 No. 06S Jun 2013, pp. 26-29