ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

K-MEAN Approach for Improving Text Document Clustering

Parashram T. Rathod [1], Prof. B. C. Melinathmath [2]

M. Tech, Dept. of CSE BLDEA's Dr.P. G. Halkatti, College of Engineering, Vijayapur, Karnataka, India[1]

Assistant Professor, BLDEA's Dr.P. G. Halkatti, College of Engineering, Vijayapur, Karnataka, India[2]

ABSTRACT: Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications.

KEYWORDS: text search, text clustering, text classification and categorization.

I. INTRODUCTION

Clustering was developed for the numeric, but as technology grows the digitalization of the text document increased so, searching of the text document becomes very complex. For grouping a similar characteristic data, text document clustering is used now a days. For textual cluster text document clustering is application.

Clustering applications are used for information filtering and document organization. Web document clustering is used for research purpose; document clustering is divided in to two types online and offline clustering. There is different type of clustering methods, Hierarchical clustering, agglomerative clustering, K-Mean clustering, Bisecting K-Mean clustering. In this project, we have implemented K-Mean algorithm for text document clustering. We have used term frequency and sentence frequency equations for calculating the total number of words and number of words available in the sentence. To calculate the distance between two clusters we used Euclidian distance algorithm. The future work is projected to provide proficient text classification technique and cluster analysis technique. In text clustering we can make cluster of similar characteristic items. The similarity can be checked by using the text categorization, but there may be difference between two clusters.

In text clustering the logically similar objects are stored together. We reduce the number of disk to increase the efficiency in the data sets. In clustering the similar property items are stored in one group and other in another group but the whole group is accessible. Clusters are not predefined which means that result of clusters are not known before the execution of clustering algorithm. These clusters are extracted from the dataset by grouping the objects in it. For some algorithms, number of desired clusters is supplied to the algorithm, whereas some others determine the number of groups themselves for the best clustering result. Clustering of a dataset gives information on both the overall dataset and characteristics of objects in it.

II. RELATED WORK

In [1], Michael Steinbach George Karypis Vipin Kumar (2001) projected two main techniques for clustering that are agglomerative hierarchical clustering and K-Mean clustering. It has a drawback that is quadratic time complexity.In [2], Thangamani.M and P.Thangaraj et al., proposed a method for improving the feature selection mechanism and integrated semantic clustering. It has advantages that it produces more accurate result for even it reduces the cube size.
In [3], Manning and Raghavan proposed Hierarchical clustering algorithms that are either top-down or bottom-up. This algorithm has a drawback; it shows the result only one cluster which contains all documents. In [4], Han and Kamber proposed that clustering has its roots in many areas, including data mining, statistics, biology, and machine

DOI: 10.15680/IJIRCCE.2015. 0309112

learning. It has an advantages, it is useful in exploratory data analysis, grouping, decision making, data mining, information retrieval, image segmentation, and pattern classification. In [5], Samah Fodeh, Bill Punch and Pang-Ning (2010) developed the model which is based on synonymous and poly-synonymous nouns that are available in the document group. Advantage in this algorithm is that, it reduces the properties required for document clustering. In [6][7], Zamir and Etzioni proposed the method Suffix Tree Clustering, which lies between an information search engine and the user. It has some advantages; it provides the structure to the stream of data which analysis the query result. In [8], Rekha Baghel has written a common method for document clustering. Advantage of this algorithm is the FCDC (Frequent Concepts based document clustering) which works on frequent concepts than on frequent objects used. In [9], Berry Michael W proposed the method for automatically extraction of information from dissimilar resources. Drawback for this algorithm is pushing all the objects that presently are not applicable to the requirements in order to get the significant information. In [10], K. Raja and others proposed system that designed to identify the semantic relations using the ontology. The ontology is used represent the term and concept relationship. Drawback for this algorithm is the cube size is very high and accuracy is low. In [11], Sun Park and others proposed method NMF (non negative matrix factorization) which uses weighted semantic features to find out the cluster similarity. It has an advantage that group documents easily also improves the clusters quality.

III. PROPOSED ALGORITHM

Input: domain name, data samples
Output: categorized data list
Process:
1. Read the text document
2. Pre-process the text in order to remove unnecessary tag and text
a. Calculate Term frequency

$$Tf = \sum_{i=0}^{n} \frac{Word_t}{total\ words\ in\ document}$$

b. Calculate Sentence formation frequency.

$$Sf = \sum_{i=0}^{n} \frac{Word_t}{total\ sentence}$$

c. Create name value pair using array list
3. Sort array list
4. Find out the distance between domain keywords and input text using K-Mean clustering.

$$d(x,y) - \sum_{i=0}^{n} \sqrt{(x_i - y_i)^2}$$

5. The closest distance shows the document's category

IV. SIMULATION RESULTS

4.1 Data collection

Tf–idf stands for term frequency–inverse document frequency; it is a numerical statistic which shows the important a word in the collection of documents. It also used for weighting factor for information retrieval and text mining.

In Equation 2.7, suppose that we have a set of English text documents, we want to determine which documents are most relevant for query "Engineer makes the world ". A simple way to start eliminating documents which does not contain these four words "Engineer", "makes", ”the”, and "world", but still this contains many documents. To differentiate them we might count the number of time the term contain in the document and sum them all together. The number of time the term occurs in the document called the term frequency.

Here the term “the” is most common, which incorrectly emphasize documents because of the term “the” is more common. Instead of giving the important to the more meaningful term "Engineer", ”makes”, and "world". The term "the" is not a good keyword to differentiate the relevant and non-relevant documents. So that an inverse document

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

frequency which gives less important to the more weighted terms that occurs more frequently in the documents and increase the weight of the term which occurs less frequently in the documents. In equation 2.8, SF stands for sentence frequency and is calculate with respect to total words and total sentences in the documents.

In the equation 2.9, d(x, y) is used to find the total distances in between the point x and y. In mathematics, the Euclidean distance is the straight line distance between two points in Euclidean space. This Euclidean space called a metric space. This norm is also called the Euclidean norm.

4.2 Parameter Setting

Input parameters are as follows

- **Select the documents to collect data from:**

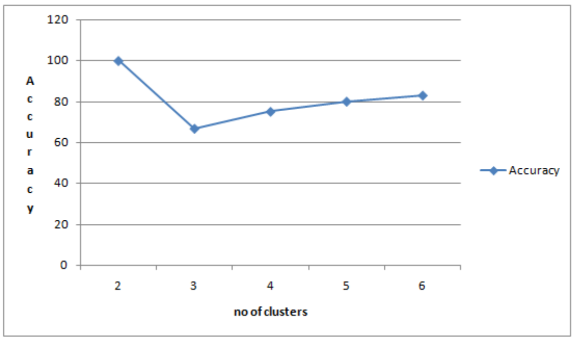
These are the documents from which the normal data or medical data or educational data collected. The user defines the number of clusters that are by the proposed system.

- **Number of Clusters to display:**

This defines the how many number of clusters defined by system, number of documents to be displayed in the final output.

4.3 Evaluation Measures

As shown in figure 4.1, 4.2 and table 4.1,4.2 the accuracy level is in between 66.66% and 83% for clustering text document. In this system number of clusters defined by user and total number of documents are varies and it depends upon user. As number of document varies the accuracy level changes also when number of clusters varies the accuracy changes.



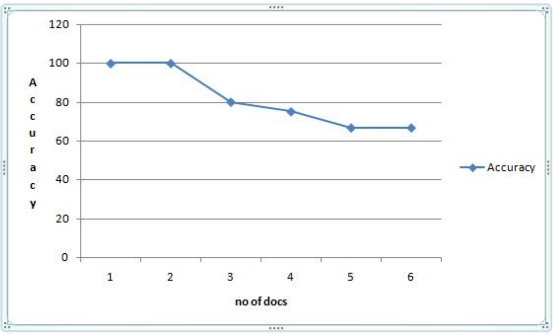
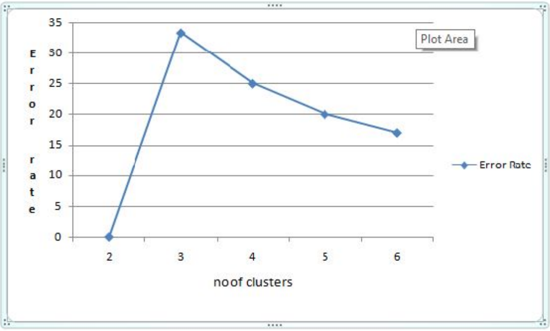


Figure
4.1: Accuracy versus number of clusters Figure 4.2: Accuracy versus number of documents



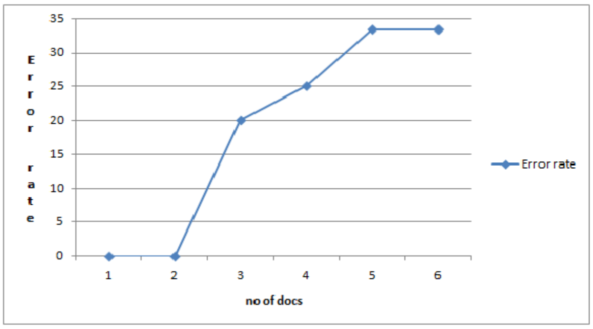


Figure 4.3: Error rate versus number of clusters Figure 4.4: Error rate versus number of documents

V. Conclusion and Future Work

The purpose of this project is to develop K-means algorithm, which can perform text document clustering on the given text data in less time. This project we is experimented with educational data.

The projected work is used for text classification and clustering technique analysis, so that the K-MEAN is used for text clustering which categorized the text in dissimilar group. In text labeling and mining the matter is of resource utilization, so that feature removal technique is used to remove the text.

This system is developed and tested by using Microsoft visual studio 2012 and performance factor shown in the figure 4.1 and figure 4.2.

REFERENCES

1. Michael Steinbach, George Karypis and Vipin Kumar(2001), ―A Comparison of Document Clustering Techniques‖, Department of Computer Science and Engineering, University of Minnesota, Technical Report 00- 034
2. Thangamani. M and P. Thangaraj," integrated clustering and feature selection scheme fortextdocuments",J.Comput.Sci.,6:536.541,DOL:10.3844/jcssp.2010.536.541,URL:http://www.thescipub.com/abstract/10.3844/jcssp.20 10.536.54
3. Manning, C. D., Raghavan, P., & Schu¨tze, H. (2008). Introduction to information retrieval. Cambridge: Cambridge University Press.H.H. Crokell, "Specialization and International Competitiveness," in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)
4. Han and Kamber, Data Mining Concepts and Techniques, Morgan Kauffman Publishers.
5. Samah Fodeh · Bill Punch · Pang-Ning Tan (2011), ―On ontology-driven document clustering using core semantic features‖, Received: 10 December 2009 / Revised: 6 September 2010 / Accepted: 26 November 2010, Springer-Verlag London Limited 2011
6. Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In Research and Development in Information Retrieval, pages 46–54, 1998.
7. Oren Zamir and Oren Etzioni. Grouper: a dynamic clustering interface to Web search results. Computer Networks (Amsterdam, Netherlands: 1999), 31(11–16):1361– 1374, 1999.
8. Rekha Baghel and Dr. Renu Dhir (2010), ―A Frequent Concepts Based Document Clustering Algorithm, International journal of Computer Applications (0975-8887), Volume 4-No.5,July 2010.
9. Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.
10. Prof. K. Raja , C. Prakash Narayanan, "Clustering Technique with Feature Selection for Text Documents", Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010 6 February 2010, Chennai, India.
11. Sun Park, Dong Un An, Choi Im Cheon, "Document Clustering Method Using Weighted Semantic Features and Cluster Similarity," digitel, pp.185- 187, 2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, 2010.