International Journal of Contemporary Technology and Management
Volume VI, Issue IX, September - 2017 ISSN (Online) 2278-8034

Text Document clustering using K-means clustering and word graph

Nidhi Bhandari[#1], Abhilasha sharma[*2]
[#]*SD Bansal College of Engineering Indore*
nidhi.bhandari1@gmail.com
abhilasha_sharma80@yahoo.co.in

Abstract – **text mining and content analysis is respectively a classical domain of research and study. A significant amount of growth is observed in this domain for supporting various applications in real world. In this presented work a text mining approach is presented for text content analysis and their clustering. Basically the purpose of text clustering is to group the text data according to their subjective context. In this context the traditional k-means clustering is applied on text data with minor modification for improving the results of clustering. The paper also contains the experimental results, the evaluated results demonstrate the proposed technique outperform as compared to traditional k-means based text clustering approaches. Additionally the future extension of the current methodology is also discussed.**

Keywords: text clustering, text mining, k-means algorithm, performance evaluation, algorithm improvements

I. INTRODUCTION

Mining is a technique by which the required objects are extracted from the raw sources. In this context the text mining is a technique by which essential text is extracted from raw source of data or information. Additionally for performing the mining the computational algorithms are employed as a tool for recovering the outcomes. The mining technique can be depends upon the utilization of data in some application. Therefore according to the need of patterns and data requirements the data mining tools and algorithms are varying. In a broad manner the data mining technique can be supervised which supports the classification and can be unsupervised supports the clustering of data.

Classically the unsupervised techniques are used when the huge amount of data exists for filter or separating them with some constrains. Therefore the clustering approaches are works as filter for data and categorize them according to their similarity. In this presented work the text clustering approach is investigated and implemented. The proposed technique usages the traditional k-means clustering with minor modification on existing methodology of k-means clustering. The concept is based on finding the centroid using the selected optimal feature set that includes much frequent set of information relevant to the clusters. In addition of that some text feature evaluation techniques are also included for effective feature selection analysis.

This section provides the overview of the proposed work. In addition of that the initial concept of the work is also explained. In further the traditional k-means algorithm is explained and the proposed methodology is explained.

II. ALGORITHM STUDY

In literature a different nature of clustering approaches are available. Among them in partition-based cluster analysis k-means clustering is a popular algorithm [4]. In this context the algorithm leads to create k number of partitions from the initial input data. In order to cluster data in k number of clusters first need to select k random objects from the available objects as initial cluster centers. These initial clusters are the points under which the all other available data objects are clustered. After election of initial cluster centroids the distance between each object and each cluster center named as centroid is calculated. Basically the distance of the object and centroid is a measurement of dissimilarity of object. Additionally using this distance the nearer points of cluster is identified and assigned. That is the first phase of clustering in next phase the optimization of centroids are taken place during this the average of all clusters are used to establish new centroid that process is repeated until the criterion function converged. Square error criterion for clustering can be given using the following formula:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2$$

x_{ij} is the sample j of i-class, m_i is the center of i-class, n_i i is the number of samples of i-class.K-means clustering algorithm is simply described as

Input*:* N objects to be cluster $x_1, x_2, \ldots, x_n$, k number of cluster
Output*:* k clusters and sum of dissimilarity between objects

and nearest centroids;

Process:

1. Select k objects randomly as initial centroid$(m_1, m_2, \dots, m_k)$;

2. Compute distance between each object x_i and centroid m_k, Assign each object to nearest cluster, distance calculation can be performed as:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_i - m_{j1})^2}, i = 1 \dots N, j = 1 \dots k$$

$d(x_i, m_i)$is the distance between data i and cluster j.

3. Calculate mean of objects in each cluster for finding new cluster centeroid,

$$m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$$

N_i is the number of objects in cluster i;

4. Repeat 2, 3 step until function E converged, return$(m_1, m_2, \dots, m_k)$.

Table 1 k-means clustering

III. PROPOSED WORK

The proposed working model is discussed in figure 1. The given model is intended to discover the text clusters according to the traditional k-means clustering with the minor modification for scaling the performance of clustering and their accuracy. The required components of the proposed text clustering model are described as:

1. Input dataset: for any data mining and machine learning based system the data is an essential part of knowledge discovery. Therefore the initial set of data is required for process and extracts the required application centric patterns from data. In this presented work the text clustering technique is main aim of design and demonstration. Therefore a collection of different subjective data is prepared as the initial dataset for the proposed clustering model. This collection contains a number of text files which contains the different subject information.

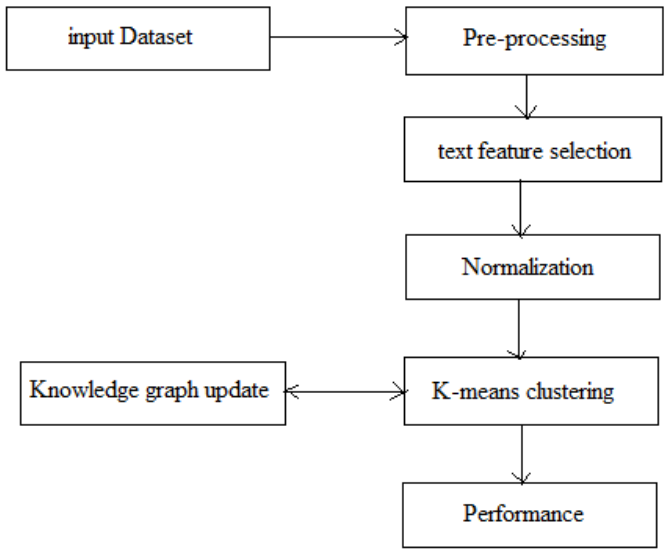

Figure 1 proposed clustering model

2. Pre-processing: pre-processing is a technique of data filtering and refinement. Using this technique the quality of learning data is improved. In this context the unwanted data or symbols are removed or filtered from the data. In this given approach the data is evaluated in two phases as:

A. removal of special characters: that is the initial phase of text processing where the special characters are removed from the input text. In order to perform this task a list of special character is prepared and each character is replaced from the entire text block.

B. removal of stop words: this phase is a second process of information filtering. In this phase the stop words (i.e. is, am, this, that, hello) similar words are removed from the input text. Basically the stop words are the additional content in text dataset which are not making valuable effort for identification of the domain of information which are need to be process or recognize using the computational algorithms.

3. Text feature computation: after removing the information or pre-processing of data the key terms from each input text files are need to be extracted. Therefore each text file is treated as the individual data object. These text file is tokenized first and for each token the term frequency is computed. The term frequency is computed using the following formula:

$$T_f = \frac{T_c}{T_w}$$

Where the is the term frequency of the individual word and count of words which is obtained in a single file finally the is the total words count in a selected file which is being processed.

This phase return the list of words and the corresponding list of word frequency. That is passed to the next phase for processing of the data.

4. Normalization: the outcome of the previous phase is produced in this phase for normalizing the feature vector. The normalization is a technique to scale the data in a similar manner. But here the text data is used thus the length of features are scaled. Because the length of all the input text files are dissimilar in size and corresponding list of extracted tokens are also not similar in all the files thus using the min-max technique the common length of features list if prepared. Thus first need to identify the maximum number of features in a file and the minimum length of features from a file, both the values are used for finding a common length of file. In order to compute the common length of feature the following formula is used:

$$F_l = \frac{max - min}{N}$$

Where is the length of feature which is need to be compute, max is the maximum number of features in a file and min is the minimum number of tokens in a file finally the N is the total number of files for analysis.

After computing the common feature length all the file of data is equalized as the computed feature length. If the number of features are larger than common length the high frequent tokens are kept and remaining tokens are removed similarly if the length of token is less then additional null tokens are appended to the feature vector. Using this technique all the files data is converted into the 2D vector which is used with the k-means algorithm in next phase.

5. k-means clustering: in this phase the 2D vector is processed using the k-means algorithm. That algorithm is works in two phases:

1. If knowledge graph is empty: before initialization of algorithm the system evaluate the knowledge graph, If the graph contains the domain specification or not, if the graph not contains anything the system randomly select the centroids from the available data objects. Then find the distance and then update the centroids. For updating the centroids the common words from all the list of files are selected and new centroids are created.

2. If knowledge graph contains the tokens: if the knowledge graph contains the domain specification the domain based token list is extracted from graph and then used as the centroid for the clustering algorithm. During the optimization phase the less number of optimization cycles the clustering fulfilling the criteria.

6. Knowledge graph update: after completing the clustering the computed centroids are updated as the cluster specification and preserved for future cluster formation. In addition of that during the single or small number of files on which clustering is required this graph directly used with their centroid based tokens for assigning the file into the cluster.

7. Performance: after clustering the performance of the proposed system is also computed in terms of accuracy. In addition of that the performance of system based on memory and time consumption is also evaluated.

IV. RESULTS ANALYSIS

This section provides the performance analysis of the implemented approach of text clustering. In addition of that the comparative performance study with respect to traditional k-means clustering is also provided.

A. Accuracy

In any data mining and machine learning algorithm the accuracy is the measurement of total amount of data accurately recognized during the testing of algorithm. The accuracy of the algorithm can be computed using the ratio of total correctly recognized data objects among total produced input to recognize.

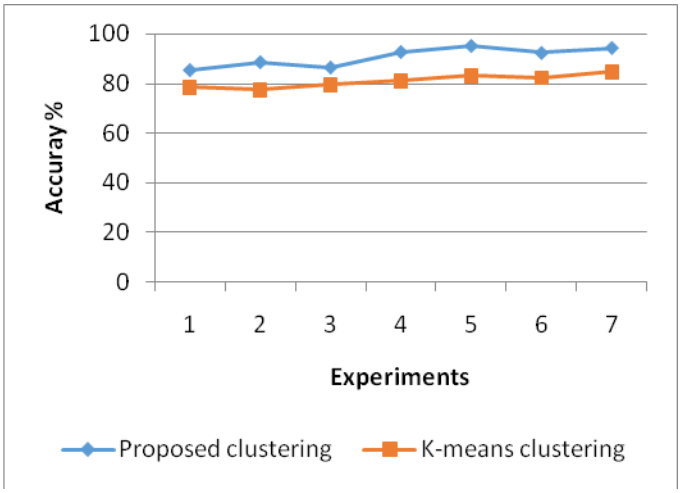

Figure 2 accuracy

The accuracy of both the algorithms namely proposed k-means clustering and traditional k-means clustering algorithm is demonstrated using figure 2. The X axis of the diagram contains the different experiments performed with the algorithms with the increasing amount of data. In addition of that the corresponding amount of correctly clustered documents in terms of percentage is given in Y axis. According to the obtained results the proposed algorithms demonstrate the superiority over the traditional k-means clustering algorithm. In addition of that the proposed algorithm demonstrates the accuracy between 80-96% and the

traditional algorithm demonstrate the outcomes between 78-86%. Thus the proposed technique is more effective than the traditional approach of clustering.

B. Time Consumption

The amount of time required to compute the clusters according to input data is termed as the time consumption or time requirement of the algorithm.

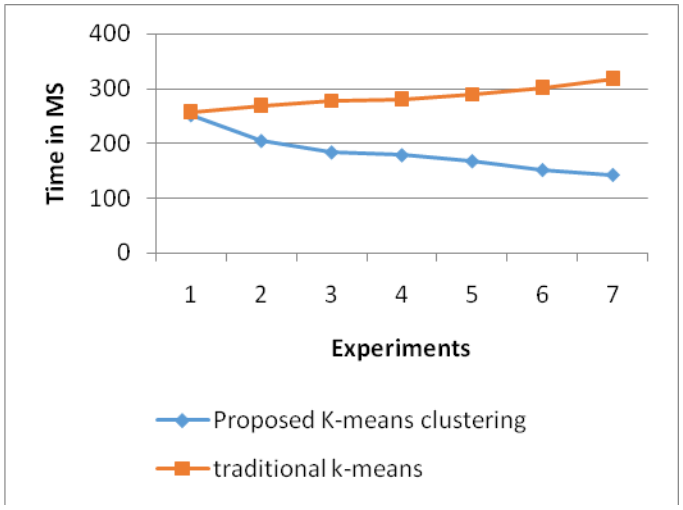

Figure 3 time requirements

The figure 3 shows the time consumption of both the algorithms for text clustering approaches. The measured time is defined here in terms of milliseconds (MS). The blue line in this diagram contains the performance of proposed algorithm and the red line shows the performance of traditional algorithm. In the similar manner the X axis of the diagram contains the different experiments performed with the algorithms and the Y axis shows the corresponding expend time in terms of milliseconds. According to the observed results the proposed technique's time consumption is decreases with the amount of data and number of experiments as compared to the traditional algorithm. In addition of that the traditional algorithm needs the similar amount of time or continuously increasing amount of time for clustering the documents. Thus the proposed technique is efficient as compared to the traditional technique of k-means clustering.

C. Memory Usages

Process needs a specific amount of main memory to hold variables and intermediate outcomes during the process execution this amount of memory is known as the space complexity or memory consumption of the algorithms. The comparative performance of both the algorithms namely traditional k-means clustering and proposed graph based clustering approach is demonstrated using figure 4. In this diagram the X axis contains the different experiments performed with the system and the Y axis shows the consumed memory in terms of KB (kilobytes). According to the observed results the proposed technique reduces the memory consumption with the amount of experiments with the same algorithm even the amount of data for clustering is increases in addition of that in traditional technique the amount of memory requirement is increases with the increasing amount of data. Thus the proposed technique enhances the resource consumption of the clustering approach.

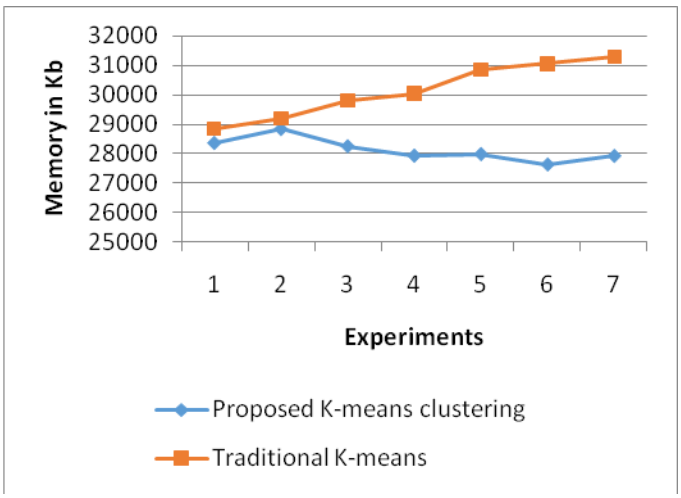

Figure 4 memory usages

V. CONCLUSION AND FUTURE WORK

The data mining and techniques enable us to analyze the complex and large amount of data automatically with placing the human efforts. In addition of that these techniques are also suitable for obtaining the patterns that are suitable for a specific kind of application. These algorithms either supervised or unsupervised in nature. The supervised algorithms are efficient and the accurate as compared to the unsupervised algorithms. Therefore in this presented work the unsupervised learning and their performance improvement is the key area of investigation and design. The proposed work is focused on improving the technique of k-means clustering based text clustering using the graph introduction in clustering. The method preserves the previous clustering knowledge in terms of graph model and utilized when the again clustering is called for the similar kind of data clustering. That technique not only helps to find optimum clusters it also helps to improve the clustering performance in terms of accurate cluster deign. Finally the evaluation of the resource consumption justifies the outcomes when it requires fewer amounts of time and memory for performing the clustering of the large amount of text files.

In near future the aim is to extend the implemented technique for introducing the semantic during the clustering operation in text files. In addition of that their applications in other similar domain mining techniques are also investigated.

REFERENCES

[1] Wang, Juntao, and Xiaolong Su. "An improved K-Means clustering algorithm" Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on IEEE, 2011.
[2] Sechelea, Andrei, et al. "Twitter data clustering and visualization", Telecommunications (ICT), 2016 23rd International Conference on, IEEE, 2016.
[3] Kinsella, Sheila, Alexandre Passant, and John G. Breslin, "Topic classification in social media using metadata from hyperlinked objects", European Conference on Information Retrieval, Springer Berlin Heidelberg, 2011.
[4] Ho-Hyun Park ,Jaehwa Park and Young-Bin Kwon, "Topic Clustering from Selected Area Papers", Indian Journal of Science and Technology, Vol 8(26), October 2015
[5] Berkhin, Pavel, "A survey of clustering data mining techniques", grouping multidimensional data, Springer Berlin Heidelberg, 2006. PP. 25-71.
[6] H. Karanikas, C. Tjortjis, and B. Theodoulidis, "An approach to text mining using information extraction," in Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, 2000.
[7] Umajancy. S, Dr. Antony SelvadossThanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013
[8] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009
[9] Dr. S. Vijayarani,Ms. J. Ilamathi and Ms.Nithya, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks,, Volume 5(1),7-16
[10] R. Feldman and I. Dagan, "KDT - knowledge discovery in texts", In Proc. of the First International Conference on Knowledge Discovery (KDD), pages 112–117, 1995.