International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

Improved Clustering of Documents using K-means Algorithm

Merlin Jacob
Department of Computer Science and Engineering
Caarmel Engineering College, Perunadu
Pathanamthitta, Kerala
M G University, Kottayam

Anina John
Assistant Professor in CSE
Caarmel Engineering College, Perunadu
Pathanamthitta, Kerala
M G University, Kottayam

***Abstract*— With the huge upsurge in information, it has become difficult to gather relevant information within the limited time. Hence clustering methods are introduced to ease the task of gathering the relevant information in a cluster. Efficiency of clustering therefore becomes one of the crucial requirements to be met by the clustering methods. There are several methods and algorithms have been introduced. Hierarchical clustering is often portrayed as the better quality clustering approach, but it is limited because of its time complexity. In contrast, K-means and its variants have a time complexity which is linear in the number of documents. A clustering method based on the hidden semantics within the documents is proposed here for better results. The proposed method extracts features from the web documents using conditional random fields and builds a linguistic topological space based on the associations of features. The features that are used this method are TF (Term Frequency) and IDF (Inverse Document Frequency). Both TF and IDF values are best in reflecting the importance of the document in the given context. Then the documents are clustered based on the K-means clustering after finding the topics in the documents using these features. The advantage of K-means method is that it produces tighter clusters than hierarchical clustering, especially if the clusters are globular.**

Keywords*— Document clustering, TF, IDF, K-means, cosine similarity, heirarchichal clustering.

I. INTRODUCTION

We are currently standing in the age of information where everyday huge amount of data is accumulated in the digital form. Therefore the problem of how to interpret or gather or analyze the information in the hand is always there. Clearly we need powerful measures to overcome this problem and they should be able to adapt to the varying situations quickly without fail. Clustering analysis is the sub-field of artificial intelligence that is brought to solve this problem of information age. Document clustering [1], [2], [11] is a technique that is used in grouping of documents into relevant clusters or groups based on some metrics. For a good clustering technique documents lies within the same cluster or group should be similar in nature as possible and two different documents in two different clusters should also be different from each other. By using this clustering technique we can reduce the amount of work we need to do when searching or browsing for knowledge. Other than its use in search engines it can also be used in other fields like market segmentation, medical analysis, text mining, information retrieval etc [5], [6].

Document clustering is a sub area of data clustering which includes concepts from information retrieval, natural language processing, and machine learning [3]. The main aim of document clustering method is find out natural groupings of documents from a given collection of documents. It is a total different concept from classification. In classification the number of classes is known a priori [6] and then the documents are assigned to them. Conversely in a clustering technique we know nothing about the class in advance. A clustering method should be able to analyze the property of the given documents by using appropriate features and a document model. Document clustering methods usually represent a document as a vector of its selected features. Clustering algorithms are evaluated based on different metrics and the choice of metrics is purely depends on the application area of that method.

Clustering is an active research subject in the fields of statistics, pattern recognition and machine learning. Even though we are using the clustering techniques in many fields, the main use of clustering techniques stands in the field of data mining. In data mining the clustering technique handles very large amount of datasets and their properties to cluster them properly. This adds very large complications to the clustering technique to be employed. Thus the main goal of such algorithms is to minimize the computational overheard by creating more accurate clusters. There are many kinds of techniques are proposed for achieving these desired properties. The two main algorithms that are used in clustering are Hierarchical clustering and K-means clustering techniques. Hierarchical clustering is slower than K-means and sometimes combination of these two also used for good results.

Hierarchical clustering techniques produce a nested sequence of partitions or a cluster of hierarchy or tree of clusters. This structure is also called as dendrogram. In this structure every node has child and sibling clusters. The main advantages of hierarchical clustering are their flexibility and ease of handling any forms of similarity [4]. But they suffer from vagueness in the termination criteria. K-means [12] is a partitioning relocation clustering method which divides data into several subsets. When we are using K-means we are using a centroid which is the mean value of all points within the cluster. This centroid represents the cluster formed and this helps the K-means methods to produce clusters in a faster rate than hierarchical methods.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

To identify and discriminate the correct topics in a collection of documents, the combinations of features and their co-occurring relationships are the clue, and the possibilities display how significant they will be. Thus we first need to find the underlying conceptual structure of the document for further processing. So including more phrases into this system helps in adapting them to understand the complex phrases used in the documents. After understanding the semantics within the each document we can easily cluster them based on the semantics.

The overall architecture of the proposed work is given in the Fig.1 and the different steps of the work are given in the figure.

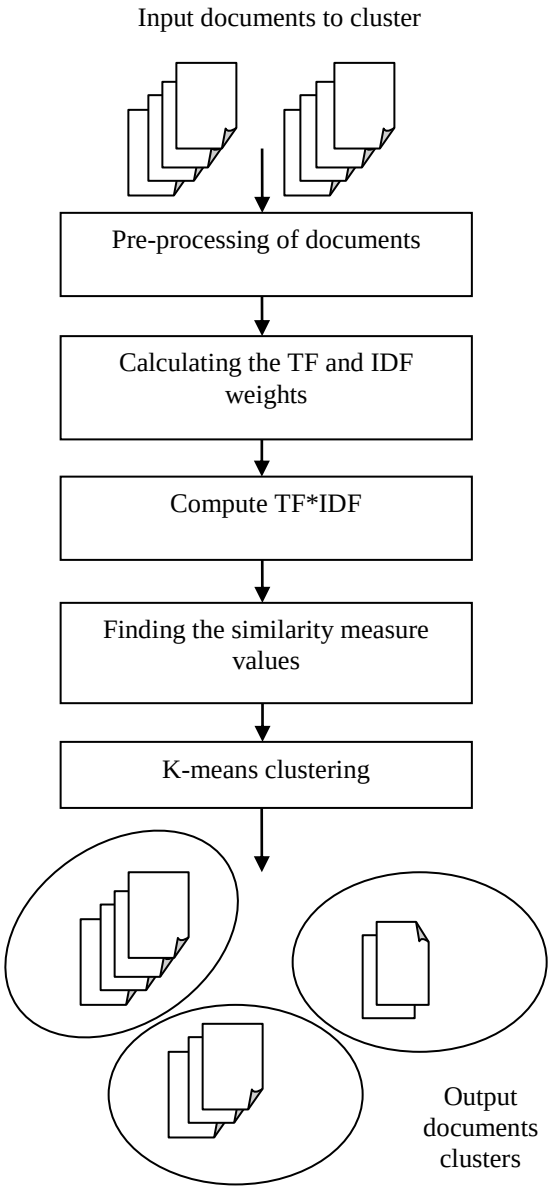

Fig. 1 System architecture

II. PROPOSED METHOD

We are currently standing in the age of information where everyday huge amount of data is accumulated in the digital form. Therefore the problem of how to interpret or gather or analyze the information in the hand is always there. Clearly we need powerful measures to overcome this problem and they should be able to adapt to the varying situations quickly without fail. Clustering analysis is the sub-field of artificial intelligence that is brought to solve this problem of information age. Document clustering [1], [2], [11] is a technique that is used in grouping of documents into relevant clusters or groups based on some metrics. For a good clustering technique documents lies within the same cluster or group should be similar in nature as possible and two different documents

A. *Pre-processing of documents*

The first step of every document clustering method is pre-processing of the input documents [6]. In pre-processing of the input collection of documents the system analyses the content of the documents and prepares them ready for vector modeling. It refines the content by word by word and removes unnecessary or redundant data from the documents. With the use of a fine pre-processing of input documents we can avoid the further complexities in the subsequent processing steps. There are many sub steps in the pre-processing of the input and they are given below:

- Filtering: In filtering [6] it special characters and punctuations that are thought to be of no meaning are removed. In the case of web documents it removes tags from the web page for further smooth processing.
- Tokenization: In tokenization [6] it splits the sentences into words as in the NLP.
- Stemming: In stemming [6] it reduces the words into their base form for the ease of the processing. For example the word “working” is reduced to its base form “work” in stemming.
- Stopword removal: Stopword [6] is a term that does not convey any particular meaning and can be ignored when modeled to vector space.
- Pruning: In pruning [6] the words with very low frequency in the entire corpus of text is avoided in order to prevent the formation of very small clusters due to their presence.

The pre-processed abstract generated from the input documents are hard to read in human understanding. But they can provide a significant improvement in retrieving relevant information from the documents. Thus the pre-processing of input documents improves the performance by considerably reducing the amount of data to be analyzed.

B. *Calculation of TFIDF*

After the pre-processing of input documents we get the abstract from the documents that are rich in relevant information. Then from the abstract we are calculating the two important values- Term frequency (TF) and inverse document frequency (IDF). These two values from the documents are used to create the vector model for the each document. The TFIDF value has the advantage of giving importance to both the term frequency and possibility of the term in the cluster over other modeling techniques. This is why most clustering algorithms are built based on the TFIDF value than the other measures.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

The TFIDF score for a term at position i in document j is defines as,

$$(TFIDF)_{ij} = (TF)_i \text{ X } (IDF)_{ij} \quad (1)$$

where $(TF)_i$ is the term frequency for the term *i* in the document *j*. $(IDF)_{ij}$ is the inverse document frequency for a term t_i is defined as,

$$(IDF)_i = \log |D| / |\{d : t_i \in d\}| \quad (2)$$

|D| denotes total number of documents and $|\{d : t_i \in d\}|$ denotes number of documents with term t_i exists.

The TDIDF scheme for weighting gives more weight to the term with less frequency and high importance in the clusters. There have been other weighting schemes also developed for more accurate clustering. The TFIDF scheme works well when the clustering is based on text corpuses. TF and IDF weights are easy to calculate than the other complex methods.

After the calculation of TFIDF weights the next step in processing is to find the similarity between documents in the input collection. Many formulas have been proposed to find out the similarity in clustering. The commonly used similarity measure is the cosine measure. Cosine measure uses the dot product of the documents to calculate the similarity. It is defines as,

$$\text{cosine}(d1, d2) = (d1 \bullet d2) / \|d1\| \|d2\| \quad (3)$$

where • represents the dot product and ||d1|| denotes the length of the vector d.

C. K-means clustering algorithm

Similarities between the documents are calculated by using the cosine measure from the vector space. Then we need to apply a clustering algorithm for clustering the documents based of the TDIDF value and the cosine similarity calculated in the previous steps.

Hierarchical clustering algorithm is always terms as a good clustering algorithm but they are limited by their quadratic time complexity. K-means and its variants have a linear time complexity which makes it an excellent choice to application where time is a crucial measure. K-means also has more run-time efficiency when compared to hierarchical clustering method. K-means method works better when the number of input documents is large.

K-means clustering works well with text documents. It first generates a set of centroid values to represent the clusters. Then for each document it calculates the value and checks to which cluster centroid the calculated value is closer. After that the document is added to that closer one and it redefines the centroid of that cluster. By repeating this process until a termination criteria the whole input can be clustered into different clusters in the final output.

K-means [8], [9] provides a faster way to create cluster from a set of random documents. It calculates the vector value for each document from the vector space and based their value new clusters are formed. Since it uses the TDIDF and cosine measure the final produced clusters are always good in terms of both intra and inter cluster similarity.

The K-Means Clustering Algorithm [7] can be defined in the following steps:

(1) Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hyper volume containing the pattern set.
(2) Assign each pattern to the closest cluster center.
(3) Recompute the cluster centers using the current cluster memberships.
(4) If a convergence criterion is not met, go to step 2.

With the use of K-means clustering algorithm it is possible to create clusters that are tight in nature. Clusters produced by this method have a higher accuracy rate than the existing methods like hierarchical clustering algorithm.

III. RESULTS

This section includes the details of the dataset used in the analysis of the proposed method and its corresponding values of the parameters.

A. Dataset

The dataset we have chosen for the analysis of this work is an English newsgroup corpus. The English dataset is called 20 Newsgroup (NG20). This 20 Newsgroup is a popular collection used as a standard dataset and it is available in the link http://people.csa1l.m1t.edu/jrenn1e/20Newsgroups/. The 20Newsgroup dataset original collection contains 19997 Usenet discussions crawled from 20 different newsgroups boards. The documents are almost evenly distributed over the newsgroups. The topics say about politics, religion, computer science, sports etc. The categories in the NG20 dataset are shown in the Table I.

B. Parameters

The target of a clustering method is to produce good quality clusters at the end of the processing. The quality of the generated clusters can be tested by various parameters. These parameters are representing the effectiveness of the total method employed.

In this work the parameters selected to check the quality of the proposed method are the total execution time, frequent item dataset and the number clusters formed.

The proposed method is analyzed against a fuzzy clustering method which employs hierarchical aggregation algorithm. This method is originally developed for the clustering of medicinal documents in a lab. This fuzzy clustering algorithm for document clustering is based on the concepts of fuzzy set theory. It calculates the membership function values for the document before the clustering to assign documents to particular clusters.

The total execution time shows the time taken by the entire method to compute the values and produce the clusters after submitting the documents to the system. The method with less execution time shows that method is faster in cluster formation with high accuracy.

The accuracy of the system shows the capability of the proposed and existing systems to create clusters with relevant similarity between them. The accuracy can be used as a great measure to compare the overall performance of the system.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

The cosine similarity of the clusters shows the average similarity of the documents within the clusters. It is a good indicator of the effectiveness of the algorithm in generating accurate clusters. The similarity of the documents within the cluster should always be greater compared to the documents lies within two different clusters.

C. *Simulation Results*

The values of the parameters used in evaluation and their corresponding graphic representation are included in this section. The three parameters used were the execution time, accuracy and average similarity in clusters. The three parameters that are used here are efficient in calculating the total performance of the system. The system is proved to be of better performance in terms of these three parameters than the existing fuzzy hierarchical algorithm and the corresponding graphic representations are also shown in this section.

TABLE I. CATEGORY DISTRIBUTION NG20

Sl. No	Topic	Number of documents
1	alt.atheism	799
2	comp.graphics	973
3	comp.os.ms-windows.misc	985
4	comp.sys.ibm.pc.hardware	982
5	comp.sys.mac.hardware	961
6	comp.windows.x	980
7	misc.forsale	972
8	rec.autos	990
9	rec.motorcycles	994
10	rec.sport.baseball	994
11	rec.sport.hockey	999
12	sci.crypt	991
13	sci.electronics	981
14	sci.med	999
15	sci.space	987
16	soc.religion.christian	997
17	talk.politics.guns	910
18	talk.politics.mideast	940
19	talk.politics.misc	755
20	talk.religion.misc	628

TABLE II. EXECUTION TIME FOR CATEGORIES IN NG20

Sl. No	Topic	Execution time in milliseconds	
		Existing System	Proposed System
1	alt.atheism	3453	735
2	comp.graphics	3000	609
3	comp.os.mswindows.misc	1000	219
4	comp.sys.ibm.pc.hardware	985	203
5	comp.sys.mac.hardware	828	172
6	comp.windows.x	235	62
7	misc.forsale	766	157
8	rec.autos	953	203
9	rec.motorcycles	2516	516
10	rec.sport.baseball	1078	343
11	rec.sport.hockey	672	250
12	sci.crypt	812	516
13	sci.electronics	1078	234
14	sci.med	1203	250
15	sci.space	7500	1422
16	soc.religion.christian	1312	297
17	talk.politics.guns	2819	594
18	talk.politics.mideast	2640	546
19	talk.politics.misc	1078	218
20	talk.religion.misc	1328	281

The Fig. 2 shows the execution time of the existing fuzzy clustering system against the proposed K-means clustering system in milliseconds. The Table II shows the execution time of the both systems when they were input the different newsgroups in the dataset NG20.

From the Table II and Fig. 2 it is clear that the execution of the proposed method is better than the existing algorithm. The execution time indicates the simplicity and fast computational capability of the proposed method when compared to the existing ones.

Fig. 3 shows the graphical representation of the system in terms of cluster similarity. Cluster similarity is calculated as the average of the similarity value of the documents in the generated final cluster. The high similarity measure shows that the documents within each cluster are tighter than the existing fuzzy clustering algorithm. In the Fig. 4 we can see that the accuracy of the system in processing data compared to the fuzzy clustering method.

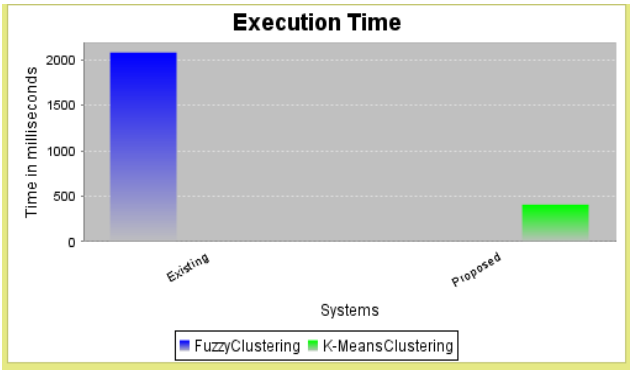

Fig. 2 Execution time of NG20 dataset

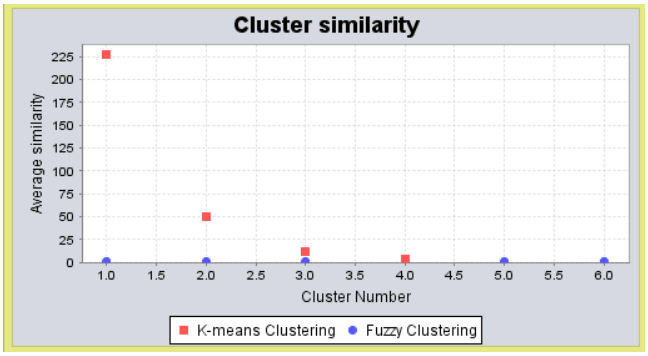

Fig. 3 Cluster similarity of NG20 dataset results

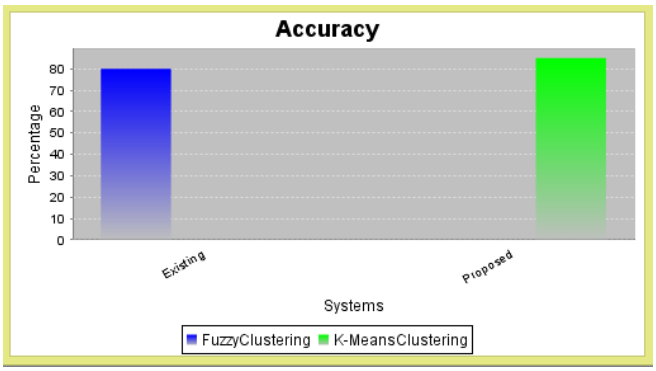

Fig. 4 Accuracy of the systems in percentage

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 07, July-2016

IV. CONCLUSION

In this paper a novel method which can be used to cluster documents into relevant clusters from a corpus of documents is introduced. The method is based on the hidden semantic structure of the documents and therefore they produce better quality clusters than the existing methods. The method uses the TF and IDF values to identify the distribution of words and their importance within that document. Then the similarity between documents is calculated for identifying the clusters. The similarity measure and TFIDF values together helps to capture the actual picture of the content within each document clearly. The proposed method proved to be capable of generating clusters with best quality and also in a faster way. The system has the advantages of high cluster similarity and reduced execution time in clustering process. With a fast and simple document clustering algorithm we can considerably reduce the time in the processing of unclassified data in business applications, search engines, marketing, educational applications, and so on. The proposed work can be extended by reducing the overall computational requirements of the system by using better measures that can easily reflect the underlying meaning within each document.

REFERENCES

[1] I-Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar, "Discovering Latent Semantics in Web Documents Using Fuzzy Clustering", IEEE transactions on fuzzy systems, vol. 23, no. 6, december 2015.

[2] Pankaj Jajoo, Document Clustering, Indian Institute of Technology Kharagpur, 2008.

[3] Christopher Issal and Magnus Ebbesson, Document Clustering, Chalmers University of Technology University of Gothenburg, Department of Computer Science and Engineering Göteborg, Sweden, August 2010.

[4] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques," Technical Report #00-034 (2000), Department of Computer Science and Egineering, University of Minnesota.

[5] Jacob Kogan, Charles Nicholas and Marc Teboulle, Clustering Large and High Dimensional data [Online]. Available: http://www.cs.umbc.edu/~nicholas/clustering

[6] Nicholas O. Andrews and Edward A. Fox, Recent Developments in Document Clustering, Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, October 16, 2007.

[7] A.K. Jain M.N. Murty And P.J. Flynn, "Data Clustering: A Review," in ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[8] Chih-tang chang, jim z. C. Lai and mu-der jeng1. A fuzzy k-means clustering algorithm using cluster, Center Displacement, Journal of information science and engineering 27, 995-1009 (2011).

[9] Md. Khalid Imam Rahmani, Naina Pal and Kamiya Arora , Clustering of Image Data Using K-Means and Fuzzy K-Means, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 7, 2014.

[10] B.S.Vamsi Krishna, P.Satheesh and Suneel Kumar R, Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-6, August 2012.

[11] Jun Tang, Improved K-means Clustering Algorithm Based on User Tag, Journal of Convergence Information Technology Volume 5, Number 10. December 2010.

[12] Rakesh Chandra Balabantaray, Chandrali Sarma and Monica Jha, Document Clustering using K-Means and K-Medoids, International Journal of Knowledge Based Computer System Volume 1 Issue 1 June 2013 [Online]. Available: http://www.publishingindia.com.

[13] S. Lawrence and C. L. Giles, "Searching the world wide web," Science, vol. 280, no. 5360, pp. 98–100, 1998.

[14] R. Kosala and H. Blockeel, "Web mining research: A survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1–15, 2000.

[15] O. Zamir and O. Etzioni, "Web document clustering: a feasibility demonstration," in Proc. 19th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval,1998, pp. 46–54.

[16] F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," in Proc. 20th Int. Conf. Comput. Linguistics, 2004, pp. 562–568.

[17] O. J. Dunn, "Multiple comparisons among means," J. Amer. Statist. Assoc.,vol. 56, pp. 52–64, 1961.

[18] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. fifth Berkeley Symp. Math. Statistics Probability, Berkeley, CA, USA: Univ. California Press, 1967, vol. 1, pp. 281–297.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)