International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 15 (2015) pp 35959-35963

A Novel Document Clustering Algorithm by Fusing Bisecting k-means and UPGMA

G. Loshma,
Research Scholar,
Jawaharlal Nehru Technological University,
Hyderabad

Dr. Nagaratna P Hedge,
Professor,
Vasavi College of Engineering
Hyderabad

Abstract- Today's world thrives with data and the data needs to be organised properly. When the data is organised, then data retrieval can be achieved in a matter of seconds. The speed of data retrieval can be boosted up by clustering related documents together. The proposed work exploits the fusion of bisecting k-means and the UPGMA algorithm, to arrive at high quality clusters. Bisecting k-means algorithm is utilized to generate cluster centroids and UPGMA refines the outcome of bisecting k-means algorithm. This refinement improves the quality of the cluster. At last, the clusters are labelled by taking the degree of recurrence of terms in all the documents of the cluster into account. The experimental results of the proposed work are satisfactory and are evident through the comparative analysis.

Keywords: Text document clustering, cluster labelling, bisecting k-means, UPGMA.

Introduction

Today's world revolves around data, with the advent of internet technologies. The internet handles several petabytes of data, day by day. Data growth is uncontrollable and thus, effective management of data is an absolute necessity. Mostly, the data is in the textual format and so a mechanism is needed to organize the data. The effective organization of data paves way for faster data retrieval. Clustering is an efficient way to boost up the data retrieval process.

Clustering is the process of grouping interrelated data together. The objective of document clustering is to assemble interrelated documents together. The documents within a cluster have more similarity, whereas documents present in different clusters show minimum degree of similarity. Clustering is mainly utilized in the areas of data mining, content based information retrieval, pattern recognition, web query search and so on.

Clustering algorithms can be categorized into the following types. They are partitional clustering, hierarchical clustering, density, grid, model, frequent pattern and constraint based clustering. Partitional clustering algorithms tend to break up the data into clusters and the documents are distributed among different clusters, with respect to an objective function [1]. The best example for this kind of algorithm is k-means and its variants.

Hierarchical clustering organizes the documents in the form of tree. The hierarchical algorithms can be designed in two ways and they are agglomerative and divisive. Agglomerative algorithms consider each document as a cluster and then combine the related documents together. This process continues until all the clusters are grouped as a whole. Thus, agglomerative algorithms follow the principle of bottom-up approach. Divisive algorithm is just the opposite of the agglomerative algorithms and they follow the top-down approach. Some of the popular hierarchical algorithms are BIRCH [2], ROCK [3], Chameleon [4] and UPGMA.

Density based clustering algorithms tend to cluster documents in random shapes. Clustering is achieved by taking the number of documents into consideration, which is density. Examples for density based clustering algorithms include DBSCAN [5], OPTICS [6] and DENCLUE.

Grid based clustering algorithms group documents by utilizing the grid structure. The grid based methods are faster. Statistical Information Grid Approach (STING) is the example of grid based clustering algorithms.

Model based clustering algorithms employ a model for every cluster and the relevance of the data is measured with respect to the model. Self-Organizing Map and COBWEB are the representatives of model based clustering algorithms.

Frequent pattern based clustering algorithms rely on dimension and the pattern is created, so as to achieve clustering. The best example for this type of algorithm is pCluster. Finally, the constraint based clustering algorithms cluster documents with respect to the specified constraints [7].

The proposed clustering algorithm is the combination of bisecting k-means and UPGMA algorithm. Bisecting k-means algorithm follows top-down approach; on the other hand, UPGMA follows bottom-up approach. As this combination takes two different approached into consideration, the clustering results will be of high quality. This work is compartmentalized into three major steps and they are document pre-processing, document clustering and labelling.

Document pre-processing is the preparatory process, which makes the documents suitable for further processing by weeding out unnecessary objects. Document clustering is the most important step in which the related documents are grouped together. Finally, cluster labelling intends to provide a name or label to the cluster to make it meaningful.

The remaining sections of the paper are organized as follows. Review of literature is presented in section 2. Section 3 is loaded with the proposed methodology. The

performance of the proposed work is analysed in section 4. Finally, the concluding remarks are presented in section 5.

Basic Algorithms

This section discusses the fundamental algorithms, which forms the underlying base of the proposed work.

A. k-means algorithm

K-means algorithm is the popular clustering algorithm, and this term was first introduced by James Macqueen in the year of 1967 [8]. The standard algorithm was presented by Stuart Lloyd in the year 1957. K-means algorithm is easy to implement and thus many clustering problems employed k-means algorithm.

This algorithm consumes minimal time for execution [9-12]. The pitfall of this algorithm is its dependency on cluster point [1,13-15]. The steps involved in k-means algorithm are presented below.

Choose k initial centre points;
Allocate all the points to the nearest centre point;
Recalculate the centre point of every cluster;
Repeat steps 2 and 3 until the centre point remains the same;

The initial centre points are needed to be chosen and then all the points are allotted to the nearest centre point. The centre point of all the clusters is calculated again and this step is repeated until there is no change in the centre points.

B. Bisecting k-means algorithm

Bisecting k-means algorithm is the improved version of k-means algorithm. This algorithm iterates by selecting a cluster and follows a principle to divide the cluster. This process gets over as soon as the required count of clusters is attained or when the whole hierarchical tree is formed. The standard bisecting k-means algorithm is presented below.

Input: Dataset, Iteration count (ic), required clusters;
Output: K clusters
1. Decompose a random cluster;
2. Compute bi-clusters;
3. Repeat step 2 until ic;
4. Consider one of the bi-cluster that generates a high quality cluster;

The clustering quality of this algorithm depends on the selected stopping criteria. This can be achieved by splitting the largest cluster and then to bisect the cluster by taking the cluster centroid into account.

C. UPGMA algorithm

UPGMA algorithm is the Unweighted Pair Group Method with Arithmetic Mean algorithm, which follows the bottom-up approach. This algorithm tends to construct a dendrogram by clubbing the two nearer clusters. The clustering process is achieved by the exploitation of distance or similarity matrix.

Proposed Methodology

A text document clustering scheme includes a document representation model, feasible similarity measure, algorithm for clustering and evaluation of clusters [16,17]. The proposed work follows the aforementioned statement and has got six important phases. They are attribute initialization, documents pre-processing, text document representation model, similarity measure, application of clustering algorithm and cluster labelling. All these phases are explained in the following subsections.

A. Attribute initialization

In this step, certain attributes such as iteration limit or the time bound can be fixed. As the proposed algorithm tends to label clusters, the threshold for terms can be fixed at this phase.

B. Document pre-processing

Document pre-processing is an important step that aims at eliminating least important terms. This makes sense that the terms which do not mean a lot to the documents are thrown out. For instance, articles, prepositions, conjunction, pronouns and reflexive pronouns have no meaning of their own and thus these terms can be eliminated from the documents. These terms are called as stop words and some of them are listed in table 1.

Table 1: List of Stop words

List of stop words			
A	an	The	About
Above	Across	Afore	After
against	Along	Aside	Beside
Among	Except	Include	In
Out	Despite	During	Below
Beyond	From	Into	Unto
To	Out	Until	Till
With	Than	Through	Upon
And	But	Because	For
So	Or	Yet	I
You	My	Me	He
She	Who	Myself	herself

All the above mentioned terms make no meaning, unless they are used with proper subject and object. Thus, these terms are removed.

C. Text document representation model

In literature, several document representation models are present. Vector Space Model (VSM) is one of the most popular document representation models, which treats textual documents as vectors. Every document collection possesses numerous documents and every document contains abundant words or terms. This can be stated by the following equations.

$$Doc_{set} = \{Doc_1, Doc_2, Doc_3, \dots, Doc_n\} \quad (1)$$

$$Doc_a = \{Tm_1, Tm_2, Tm_3, \dots, Tm_i\} \quad (2)$$

where Doc_{set} is the set of documents that holds several numbers of documents. Every document encompasses numerous words, as given by the eqn. 2.

This step is followed by the computation of weights of terms being present in the document. This step determines the degree of relationship between the term and the document.

$$Doc_{wt} = \{Tm_{wt1}, Tm_{wt2}, Tm_{wt3}, \dots, Tm_{wti}\} \quad (3)$$

The weight of the term present in the document is calculated with respect to the degree of recurrence of the term and it is computed by eqn.4.

$$Wt = k_{in} * IDF \quad (4)$$

$$IDF = log\left(\frac{k}{k_n}\right) \quad (5)$$

k_{in} denotes the degree of recurrence of the term n in i^{th} document. IDF is the Inverse Document Frequency, which symbolizes the degree of recurrence of the term n from the group of documents. Thus, the degree of association between the terms and the documents is computed by this step.

D. Similarity measure

Usually, the similarity measure computation precedes the clustering process. Similarity measure is the deciding authority, which groups correlated entities collectively. The degree of relationship between different documents is computed by the similarity measure. In literature, numerous similarity measures are available. Some of the famous similarity measures are the Euclidean distance, cosine similarity, pearson correlation coefficient and Jaccard distance [18].

This paper utilizes cosine similarity distance, as it found many applications in the area of text mining. The cosine similarity is calculated by the below given equations.

$$Sim(Doc_a, Doc_b) = cosine(Doc_a, Doc_b) \quad (6)$$

$$cosine(Doc_a, Doc_b) = \frac{\sum_{i=1}^{j} Tm_{wt(ai)} Tm_{wt(bi)}}{\sqrt{\sum_{i=1}^{j} Tm^2_{wt(ai)} . Tm^2_{wt(bi)}}}, a \quad (7)$$

The above given equation yields the result ranging from 0 to 1, which determines the relationship between the documents.

E. Clustering algorithm

This work exploits the fusion of bisecting k-means algorithm and UPGMA algorithm. Thus, the proposed work inherits the merits of both the algorithms. Bisecting k-means algorithm follows top-down approach, whereas UPGMA algorithm utilizes the bottom-up approach. This paves way for the generation of refined clustering results. The algorithm is presented below.

1. Initialize the necessary attributes;

2. Pre-process the documents;

3. Create document clusters by bisecting k-means as in sect 2.2;

4. Compute centroids of the clusters;

5. Pass the computed cluster centroid to UPGMA algorithm;

6. Refine the cluster centroids;

7. Produce k count of clusters;

The outcome of the above presented algorithm is the refined clusters of documents. The related documents are clustered together, thus the documents present in a cluster are highly correlated. On the other hand, the documents in the different clusters show lesser degree of similarity. By this way, the text documents are clustered effectively.

F. Clustering algorithm

In this step, the degree of recurrence of all terms in all documents of every cluster is found out. This is followed by arranging the terms in descending order with respect to the degree of recurrence. The term with greatest occurrence is picked up and the cluster is labelled with that term. This is given by

$$lbl = Clt(Doc\left(Tm\left(DoR(Tm_1, Tm_2, Tm_3, \dots Tm_n)\right)\right)) \quad (8)$$

Where cluster is denoted by Clt, Doc is the documents present in the cluster, Tm is the terms present in the cluster, DoR is the degree of recurrence of all terms present in the document.

This is followed by the arrangement of terms in descending order with respect to the degree of recurrence in the entire cluster. The term which is ranked first is declared as the label for the corresponding cluster. The main objective of cluster labelling is to enhance the readability. Any user can come to a conclusion about the substance of the cluster, at a streak. Thus, the intention of this work to cluster the similar documents and to label the cluster with meaningful term is achieved, successfully.

Experimental Analysis

In this section, the performance of the proposed work is analysed and the results are compared with k-means, bisecting k-means and UPGMA algorithms. The performance analysis proves the fact the fusion of bisecting k-means and UPGMA algorithm works better. The dataset exploited for this work is fbis, which is a portion of TREC-5 collection and can be downloaded from [19]. This dataset possesses 2463 documents and 17 classes.

A. Performance evaluation

The clustered outcome of the algorithms is evaluated for performance in terms of quality. The basic performance metrics employed in text document clustering are precision rate, recall rate, F-measure, accuracy rate and misclassification rate.

Precision rate:

Precision rate is the total number of documents whose actual label is x, but misclassified with label y.

$$P_{rate} = \frac{doc_{xy}}{doc_y} \times 100 \quad (9)$$

Where doc_{xy} is the total number of documents with actual label x, but wrongly classified as y. doc_y is the documents which as correctly labelled as y. Thus, a clustering algorithm works well with greater precision rates.

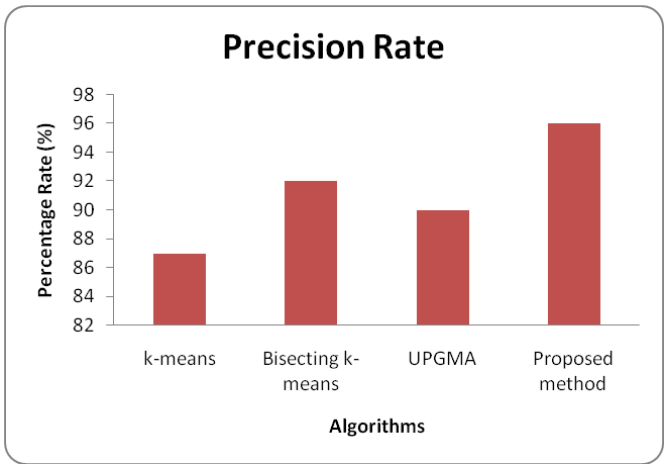

Fig 1: Precision Rate Analysis

From the experimental results, it is obvious that the proposed work shows greater precision rate with 96%. Thus, the documents are clustered in a better way.

Recall rate:

Recall rate is the total number of documents whose actual label is x, but misclassified with label y.

$$R_{rate} = \frac{doc_{xy}}{doc_x} \times 100 \quad (10)$$

Where doc_{xy} is the total number of documents with actual label x, but wrongly classified as y. doc_x is the documents which as correctly labelled as x. Thus, a clustering algorithm works well with greater recall rates.

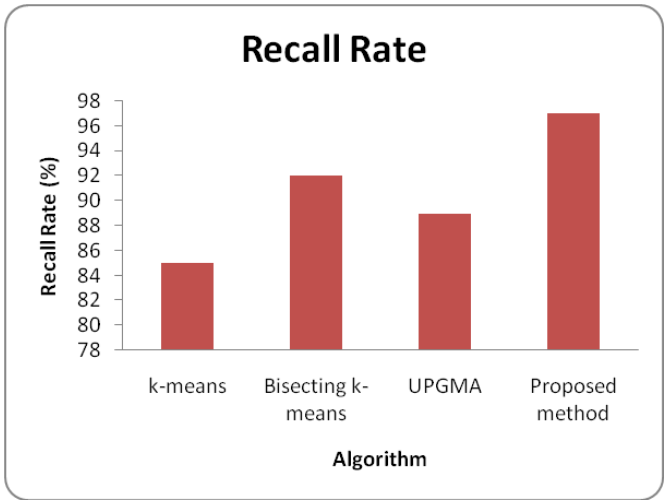

Fig 2: Recall Rate Analysis

The experimental results show that the recall rate of the proposed work is 97%, which is comparatively greater than other algorithms.

F-measure:

F-measure is computed by taking precision and recall rate into account. F-measure of a cluster and a class is given by

$$F(cls, cltr) = \frac{2 * P_{rate} * R_{rate}}{P_{rate} + R_{rate}} \times 100 \quad (11)$$

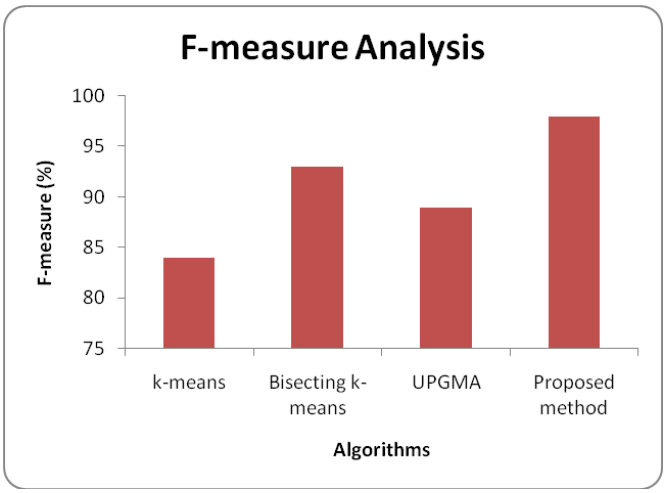

Fig 3: F-Measure Analysis

The greater the value of F-measure, the higher is the quality of the cluster. On observing the experimental results, the proposed work shows the maximum quality of cluster with 98%.

Accuracy rate:

The accuracy rate of the algorithm is determined by the sum of correctly clustered documents and the correctly rejected documents (as they are not relevant) to the total number of clustered documents.

$$acc = \frac{ccd + crd}{total\ clustered\ documents} \quad (12)$$

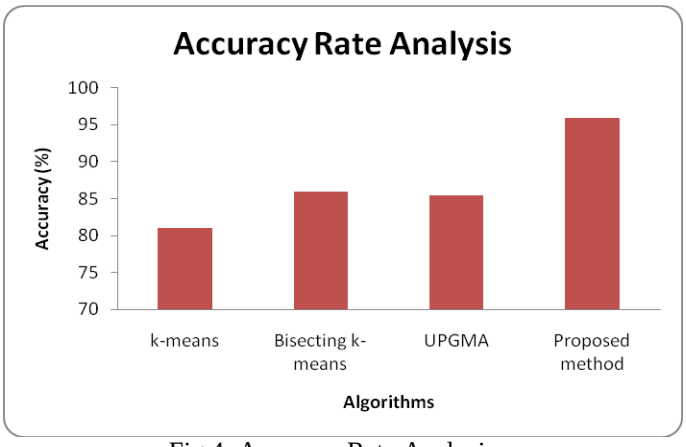

Fig 4: Accuracy Rate Analysis

The accuracy rate of the proposed work is comparatively better than other algorithms, whereby the objective of the work is fulfilled.

Misclassification rate:

Misclassification rate is the rate of wrong clustering of documents. The misclassification rate must relatively be low and is calculated by

$$mis_{rate} = 1 - acc \quad (13)$$

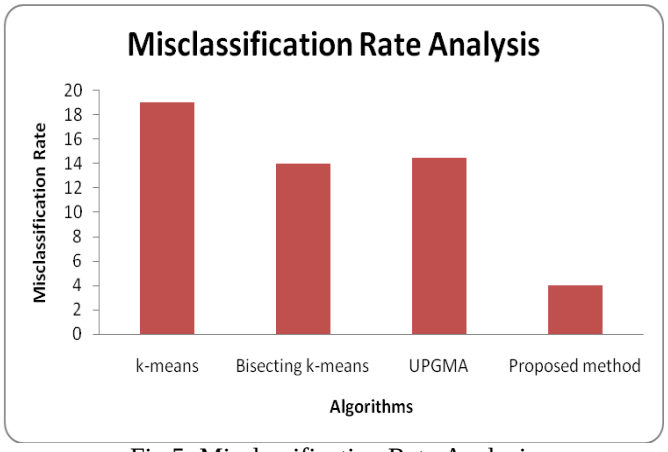

Fig 5: Misclassification Rate Analysis

Thus, the misclassification rate of the proposed work is the least, which when compared with all the other algorithms. Thus, the power of fusion of bisecting k-means and UPGMA algorithm is proven by the experimental results.

Conclusion

This work proposes to fuse the functionalities of bisecting k-means and UPGMA algorithm. The bisecting k-means algorithm is employed to find out the cluster centroid and the computed centroids are passed to the UPGMA algorithm for further refinement of clusters. The objective of any clustering algorithm is to group similar documents together and this goal has been fulfilled by the proposed algorithm. Ultimately, the readability is improved by labelling the clusters. The experimental results of the proposed work are observed to be satisfactory. In future, the quality of clusters can further be improved through semantic analysis.

References

[1] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (1999) 264–323.

[2] Zhang.T., Raghu Ramakrishnan & Livny.M.,(1996), Birch: An Efficient Data Clustering Method for very Large Databases, In Proceedings of the ACM SIGMOD international Conference on Management of Data, pp. 103-114.

[3] S. Guha, R. Rastogi, and K. Shim,(1999), ROCK: A robust clustering algorithm for categorical attributes, International Conference on Data Engineering (ICDE?99), pp. 512-521.

[4] Karypis.G, Eui-Hong Han & Kumar.V., (1999), Chameleon: A Hierarchical Clustering Algorithm using Dynamic Modelling?, IEEE Computer, Vol. 32, No.8, pp. 68-75.

[5] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).

[6] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). OPTICS: Ordering Points To Identify the Clustering Structure. ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[7] Jiawei han & Michelin Kamber ,(2010),Data mining concepts and techniques,Elsevier

[8] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on
[9] Mathematical Statistic and Probability, University of California Press, Berkley, CA, 1967, pp. 281–297.

[10] S. Lee, W. Lee, Evaluation of time complexity based on max average distance for K-means clustering, Int. J. Security Appl. 6 (2012) 449–454.

[11] C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2008.

[12] D. Reddy, P.K. Jana, Initialization for K-means clustering using voronoi diagram, Procedia Technol. 4 (2012) 395–400.

[13] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37.

[14] P. Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software Inc., 2002.

[15] J. Han, M. Kamber, A.K.H. Tung, Spatial clustering methods in data mining: a survey, in: Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001, pp. 1–29.

[16] G.H.O. Mahamed, P.E. Andries, S. Ayed, An overview of clustering methods, Intell. Data Anal. 11 (2007) 583–605.

[17] Ramiz M. Aliguliyev, Performance evaluation of density-based clustering methods, Information Sciences 179 (2009) 3583–3602

[18] Anna Huang, Similarity Measures for Text Document Clustering, NZCSRSC 2008, April 2008, Christchurch, New Zealand.

[19] http://trec.nist.gov