IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. VI (May - June 2017), PP 66-70
www.iosrjournals.org

Document Clustering Using Divisive Hierarchical Bisecting Min Max Clustering Algorithm

Prof. Vaishnavi Kamat& Prof. Terence Johnson[1]
Rudresh Chodankar, Rama Harmalkar, Gauresh Naik, Prajyot Narulkar[2]
[1](HOD & Assistant Professor,Department of Computer Engineering Agnel Institute of Technology & Design, Goa, India)
[2](B.E. Students, Department of Computer Engineering Agnel Institute of Technology & Design, Goa, India)

Abstract : *Document clustering is a process of grouping data object having similar properties. Bisecting k-means is a top down clustering approach wherein all the documents are considered as single cluster. That cluster is then partitioned into two sub-clusters using k-means clustering algorithm, so k is considered as 2. Sum of square errors (SSE) of both the clusters are calculated. The cluster which has SSE greater, that cluster is split. This process is repeated until the desired number of clusters are obtained. Divisive Hierarchical Bisecting Min–Max Clustering Algorithm is similar to bisecting k-means clustering algorithm with a slight modification. To obtain a certain number of clusters. The main cluster is divided into two clusters using Min-Max algorithm. A cluster is selected in order to split it furthers. This process is repeated until the desired number of clusters are obtained. Divisive Hierarchical Bisecting Min–Max Clustering Algorithm is similar to bisecting k-means clustering algorithm with a slight modification. To obtain a certain number of clusters. The main cluster is divided into two clusters using Min-Max algorithm. A cluster is selected in order to split it furthers. This process is repeated until desired numbers of clusters are obtained.*

Keywords: *Agglomerative clustering, Bisecting K-means, Bisecting min-max clustering, Clustering, Hierarchical clustering.*

I. Introduction

Clustering is the process of grouping similar objects based on the attributes of the object. Cluster [10] contains group of objects which are similar to each other. The need for document clustering arise due to large amount data present in unstructured manner. Document Clustering [1] is the application of Clustering of text data. Document Clustering can be used for automatic document organization, topic extraction and fast information retrieval or filtering. Agglomerative hierarchical clustering is a bottom-up approach. Initially every data object is considered as a single cluster. In each iteration merge two objects which has the maximum similarity until only one cluster remains or desired no of cluster remains. Input to the clustering technique is the BBC dataset which in turn consist of 5 classes (Business, Politics, Entertainment, Sports and Technology). There are total 2225 documents out of which 508 documents of business, 413 documents of politics, 383 documents of entertainment, 508 documents of sports and 397 documents of technology.

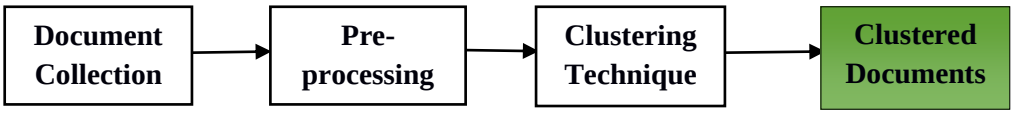


Fig. 1 Block diagram of Document processing

1. Document Collection

Collection of documents which is given as input to get desired output.

2. Pre-processing

Document Preprocessing [4] Techniques are

2.1Tokenization

The process of breaking text data into smaller units (tokens) such as word and phrases. Removing stop words and punctuation Some words are less important than others. So it is a good idea to eliminate stop words and punctuation marks before doing further analysis.

2.2 Stemming

Different tokens might carry out similar information. We can avoid calculating similar information repeatedly by reducing all tokens to its base form using porter stemming algorithm. [5]

2.3 Tf-idf(Term Frequency–Inverse Document Frequency)

It is a numerical statistic that is intended to reflect how important a word is to a document. The tf-idf value increases proportionally to the number of times a word appears in the document. [6]

2.4 PCA(Principal component analysis)

PCA is something that recognizes patterns in the data, and expressing the data in such a way that their similarity and differences are identified. It is used for analyzing data patterns
[7][8]

3. Clustering Technique

3.1 The Agglomerative Hierarchical Clustering Algorithm

Agglomerative hierarchical clustering [9] is a bottom-up approach. Initially every data object is considered as a single cluster. In each iteration merge two objects which has the maximum similarity until only one cluster remains or desired no of cluster remains.

i. Initially consider each object as a single cluster and find the similarity with respect to other objects and store it in matrix.
ii. do
iii. Combine two cluster having highest similarity.
iv. Update the matrix to find the similarity between combine cluster and rest clusters.
v. While only one cluster remains.

Result

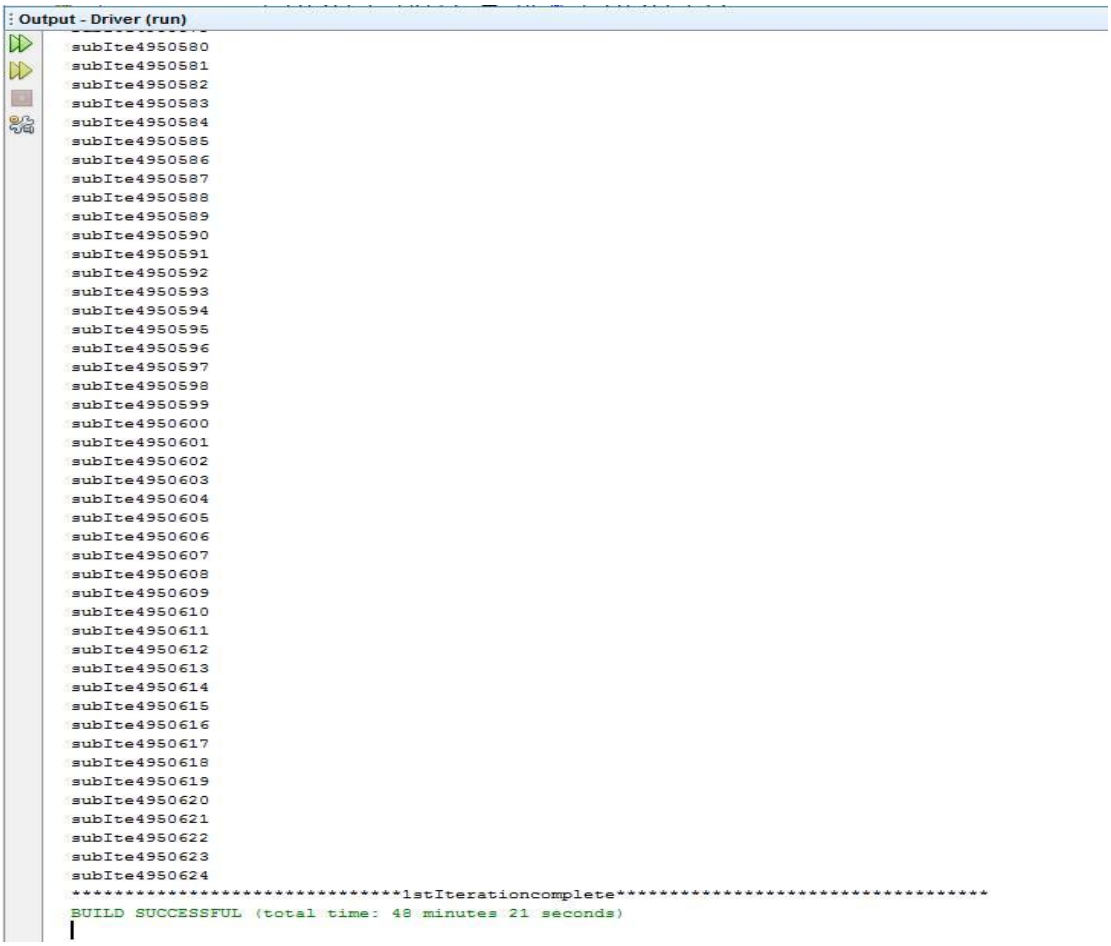


Fig 2. Result of Agglomerative clustering algorithm

The time required for the execution of agglomerative algorithm is very high. It's taking 48 minutes 21 seconds for 1 iteration. This happens because agglomerative considers each document as a single cluster and keeps on merging till a certain number of clusters are obtained.

3.2 The Bisecting K-Means clustering Algorithm

It is a hierarchical clustering method that uses basic K means clustering algorithm. It is a top-down approach. The process starts by putting all the objects in single cluster and then divides the cluster using k means i.e. K=2.Now the cluster which is having maximum sum of square error is split into 2 clusters. The process of splitting the cluster is repeated until desired no of cluster are created. [2]

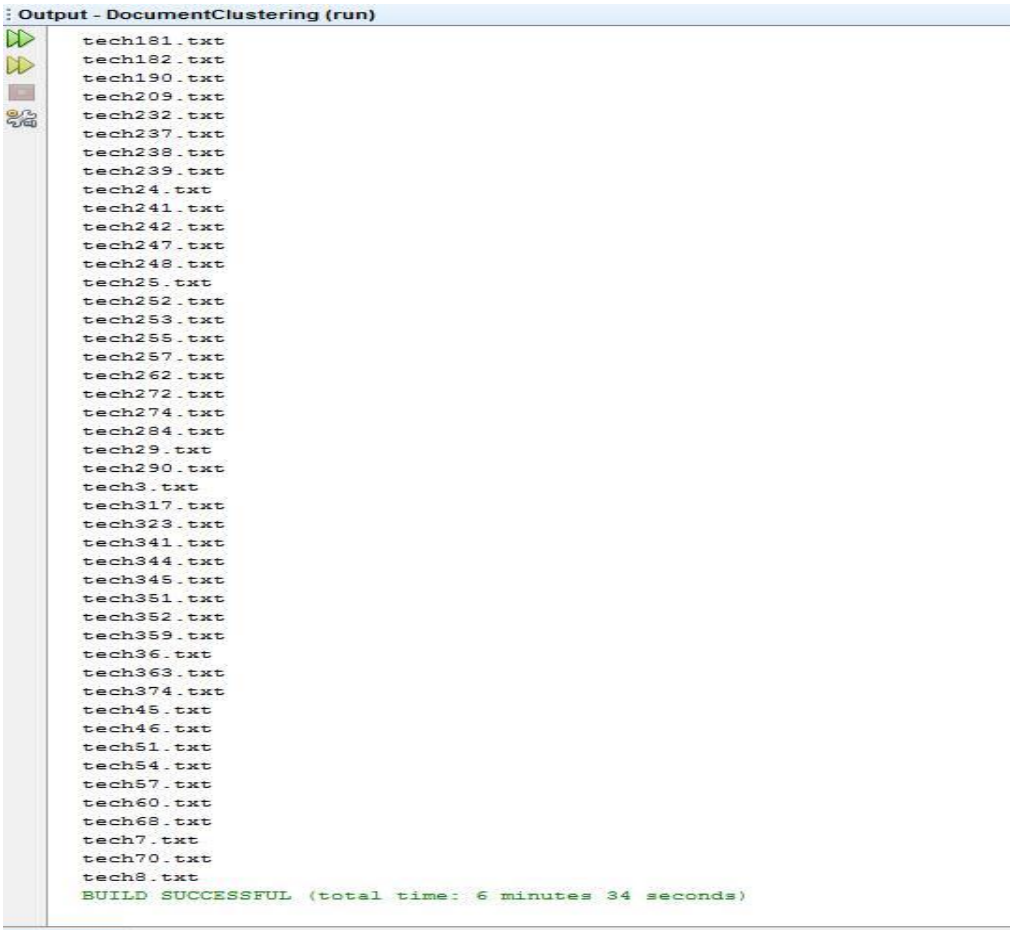
1. Initialize the list of clusters to contain all data in single cluster.
2. Initialize clustNo =1
3. while(clustNo != k)
4. Bisect the selected cluster using basic K-means
   Increment clustNo by 1
   Calculate the SSE of the clusters using

$$SSE = \sum_{i=1}^{n}(x-m)^2, \text{ where}$$

$$m = \frac{1}{n}\sum_{x} \in Ctx$$

5. If (SSE1>SSE2>….>SSEn)
   Assign cluster1 as new dataset
   Else
   Assign cluster having highest SSE as new dataset
6. End while
7. Output K clusters

Result



	Cluster1(471)	Cluster2(502)	Cluster3(415)	Cluster4(178)	Cluster5(510)
Business	23.99%	43.82%	8.67%	25.84%	17.64%
Politics	5.73%	51.39%	10.12%	11.23%	12.54%
Technology	63.26%	1.59%	2.16%	11.79%	12.15%
Entertainment	4.24%	1.99%	5.54%	33.14%	47.64%
Sports	2.76%	1.19%	73.49%	17.97%	9.80%

Fig. 3 Result of Bisecting K-means algorithm

3.3 Divisive Hierarchical Bisecting Min–Max Clustering Algorithm

This is a top-down approach. Algorithm starts by putting all the objects in single cluster then find the min point and max point from that cluster. The min point is calculated with respect to origin which is the minimum distance from origin or maximum similarity with respect to origin. From the min point max point is calculated which is maximum distance from min point or minimum similarity with respect to min point. Now 2 clusters are formed min cluster and max cluster then the SSE of each is calculated and the one having highest SSE is now treated as the new dataset to find min point and max point. This process is continued until desired no of clusters are created. [3]

INPUT: Let K=T be the number of user-specified count of clusters and a data repository
Having n data items or objects
S={S1,S2,S3,S4,…..,Sn}
Output: A set of k clusters.

- Initialize the clusters with all the points
- Initialize cluster_number =1;
- while(cluster_number != K)
- Bisect S using Min-Max clustering
  Increment cluster_number by 1
- calculate the SSE of the two clusters using the formula

$$SSE = \sum_{i=1}^{n}(x-m)^2,]]$$

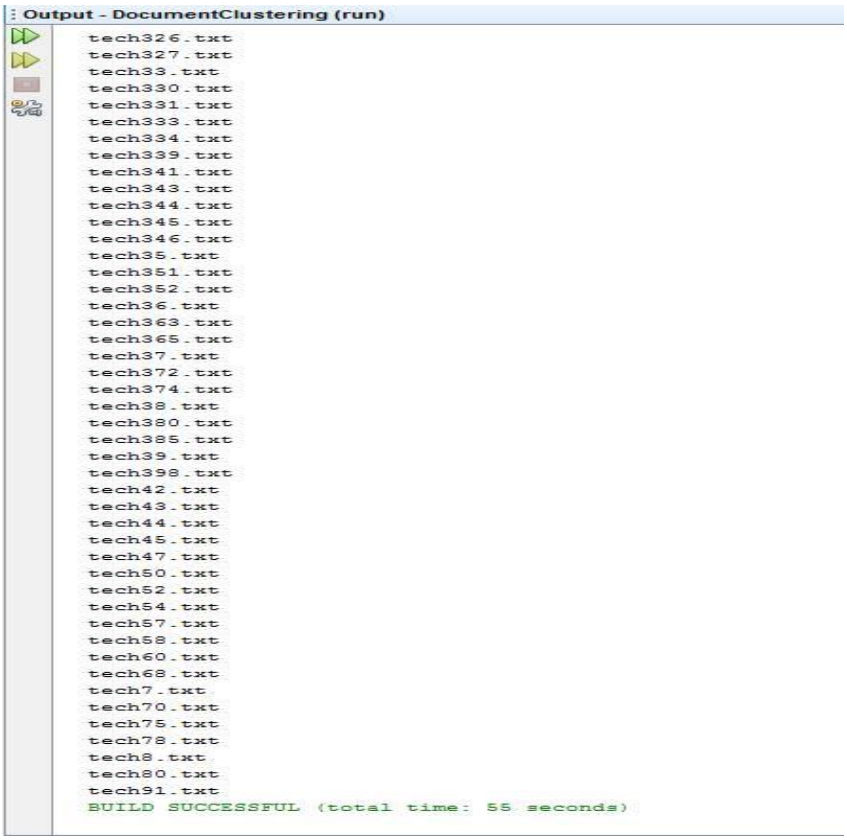
- where m is the mean of the cluster which is given as

$$m = \frac{1}{n}\sum_{x} \in Ctx$$

- If SSE1 is greater than SSE2
  then S is assigned cluster1
- else
  then S is assigned cluster2

end while
Output K clusters

II. Result



	Cluster1(209)	Cluster2(535)	Cluster3(164)	Cluster4(247)	Cluster5(1054)
Business	61.72%	12.71%	6.09%	18.62%	24.19%
Politics	7.65%	43.55%	16.46%	2.83%	12.33%
Technology	19.13%	19.43%	67.68%	8.09%	11.57%
Entertainment	1.91%	11.40%	6.09%	40.08%	19.82%
Sports	9.56%	12.89%	3.65%	30.36%	32.06%

Fig. 4 Result of Divisive Hierarchical bisecting min-max clustering algorithm

The execution time for Divisive hierarchical min max clustering algorithm is 55 seconds. Hence we can prove that Divisive hierarchical bisecting min max clustering algorithm is better than bisecting k-means and agglomerative in terms of time complexity and accuracy.

III. Conclusion

In this project, we used Agglomerative Hierarchical Clustering Algorithm, Bisecting K-Means Clustering Algorithm and Divisive Hierarchical Bisecting Min-Max Clustering Algorithm in order to cluster documents. BBC dataset is used as a dataset. As a result, it has seen that Divisive Hierarchical Bisecting Min-Max Clustering Algorithm is superior to Bisecting K-Means Clustering Algorithm which in turn is superior to Agglomerative Hierarchical Clustering Algorithm.

References

[1]. Michael Steinbach George Karypis Vipin Kumar, "A Comparison of Document Clustering Techniques" Department of Computer Science / Army HPC Research Center, University of Minnesota.

[2]. Nikita P. Katariya, Prof. M. S. Chaudhari(2015), **"**Bisecting K-means Algorithm for Text Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 5, Issue 2 February (2015)

[3]. Terence Johnson and Santosh Kumar Singh(2016), "Divisive Hierarchical Bisecting Min–Max Clustering Algorithm", Advances in Intelligent Systems and Computing Series Volume – 468, Series ISSN 2194-5357,Online ISBN 978-981-10-1675-2, DOI 10.1007/978-981-10-1675-2_57, 2016 International Conference on Data Engineering and Communication Technology-ICDECT 2016, March 10-11, lAVASA Pune, Springer Singapore, copyright 2017, copyright holder Springer Science + Business Media Singapore, pp 576-592. Clustering Algorithm"

[4]. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya*3* Assistant Professor, M. Phil Research Scholar, *"* Preprocessing Techniques for Text Mining - An Overview ", Dr.S.Vijayarani et al , International Journal of Computer Science & Communication Networks, ISSN:2249-5789 Vol 5(1),7-16

[5]. Giridhar N S, Assistant Professor, 2Prema K.V, Professor, 3N .V Subba Reddy, Professor, Department of CSE, M.I.T., Manipal University, Manipal, Karnataka, India."*A* Prospective Study of Stemming Algorithms for Web Text Mining"

[6]. R. Malathi Ravindran and Dr. Antony SelvadossThanamani(2015), "K-Means Document Clustering using Vector Space Model", Bonfring International Journal of Data Mining, ISSN 2277 - 5048 Vol. 5, No. 2 July (2015)

[7]. Lindsay I Smith (2002), "A tutorial on Principal Components Analysis" 26 February (2002)

[8]. Jon Shlens(2003), " A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS" Derivation, Discussion and Singular Value Decomposition ,Version 1

[9]. K.Sasirekha and P.Baby(2013), "Agglomerative Hierarchical Clustering Algorithm- A Review", International Journal of Scientific and Research Publications, Volume 3, Issue 3

[10]. Pang-Ning Tan, Michael Steinbach, Vipin Kumar,"Introduction to Data Mining" Pearson Education, ISBN:81-317-1472-1