Volume No:2, Issue No:11 (April-2017)
ISSN No : 2454-423X (Online)

International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

An Efficient Approach for Text Document Clustering Using Centroids and Neighbors

G Venkanna
Assistant Professor
Netaji Institute of Engineering and Technology.
Email: venkannaid@gmail.com

Abstract

Content record clustering is picking up prominence in the learning revelation field for viably exploring, perusing and arranging a lot of printed data into few important bunches. Content mining is a semi-robotized procedure of separating learning from voluminous unstructured information. A broadly considered information mining issue in the content area is clustering. Clustering is an unsupervised learning strategy that plans to discover gatherings of comparative questions in the information as for some predefined model. In this work we propose a variation strategy for discovering starting centroids. The underlying centroids are picked by utilizing most remote neighbors. For the dividing based clustering calculations generally the underlying centroids are picked arbitrarily however in the proposed strategy the underlying centroids are picked by utilizing most distant neighbors. The exactness of the groups and effectiveness of the segment construct clustering calculations depend with respect to the underlying centroids picked. In the investigation, kmeans calculation is connected and the underlying centroids for kmeans are picked by utilizing most distant neighbors. Our trial comes about demonstrates the exactness of the bunches and effectiveness of the kmeans calculation is enhanced contrasted with the conventional method for picking introductory centroids.

*Keywords:-**Text document, neighbors, kmeans, initial centroids, clusters.***

INTRODUCTION

Propelled advancements to store extensive volumes of content, on the web and on an assortment of capacity gadgets has made content reports to be accessible to the clients everywhere throughout the world with a mouse click. The occupation of orchestrating this persistently developing gathering of content archives for differing necessity of the end client is a dull and convoluted undertaking. Consequently, machine learning methods to sort out the information for speedy get to is fundamental. In the writing, there are two primary machine learning methods proposed in particular arrangement and clustering. Task of an obscure content report to a pre-characterized class is called Classification. Appointing an obscure content record by recognizing the report properties is called Clustering. Clustering a generally utilized strategy in the fields of Pattern Recognition, Mining, Data from Databases, Extracting applicable data in Information Retrieval Systems and Mining Text Data from Text records or on Web.

With always expanding number of archives on web and different vaults, the assignment of sorting out and classifying these records to the various need of the client by manual means is a muddled employment, consequently a machine learning system named clustering is exceptionally valuable. Content reports are grouped by match shrewd comparability of records with closeness measures like Cosine, Jaccard or Pearson. Best clustering outcomes are seen when covering of terms in reports is less, that is, when bunches are recognizable. Henceforth for this issue, to discover report similitude we apply connection and neighbor presented in ROCK. Interface indicates number of shared neighbors of a couple of archives. Fundamentally comparative records are called as neighbors. This work applies connections and neighbors to Bisecting K-implies clustering in recognizing seed reports in the dataset, as a heuristic measure in picking a group to be divided and as a way to locate the quantity of segments conceivable in the dataset. Our tests on continuous datasets demonstrated a huge change as far as precision with least time. One of the critical procedures of information mining, which is the unsupervised order of comparative information objects into various gatherings, is information clustering.

Report clustering or Text clustering is the association of a gathering of content archives into groups in light of comparability. It is a procedure of collection archives with comparable substance or themes into bunches to enhance both accessibility and unwavering quality of content mining applications, for example, data recovery [1], content arrangement [2], report synopsis [3], and so on. Amid record clustering we have to address the issues like characterizing the closeness of two reports, choosing the fitting number of archive bunches in a content gathering and so forth.

Document Representation

A vector space model (VSM) representation called bag of words is a simplest and widely used document representation. A vector ‘d’ is set of document terms(unique terms). In VSM the columns represent terms and row indicates document. Each row of a vector is filled with its term frequency (TF). Hence d_{tf} is given by

$$d_{tf} = (tf_1 , tf_2 , tf_3 , \ldots\ldots tf_D)$$

Where tf_i is count of occurrences of term iin d. Inverse document frequency (IDF) is the ratio of total documents(N) to the occurrence of term i in documents(df_i). IDF values are low for high frequent terms in dataset and high for less frequent terms in dataset. Log due to large dataset. Thus resulting definition of IDF is

$$IDF_i = \log \frac{n}{df_i}$$

IDF with TF is known as tf-idf weight.

$$W_{i,j} = tf_{i,j} X\ idf_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

$tf-idf$ of the document d is :

$$d_{tf-idf} == [tf_1 log(\frac{N}{df_1}), tf_2 log(\frac{N}{df_2}), .., tf_D log(\frac{N}{df_D})]$$

The centroid c_p of a cluster $Clus_p$ is given by

$$c_p = \frac{1}{|c_p|}\sum_{clus}^{doc} doc_i$$

Where |Clus$_p$| is the size of cluster Clus$_p$and doc$_i$ is a document of Clus$_p$.

SIMILARITY MEASURES

One of the essential for exact clustering is the exact meaning of the closeness between a couple of articles characterized regarding either the match wised similitude or divergence. Similitude is regularly imagined as far as difference or separation too.

Comparable records are gathered to frame an intelligent bunch in archive clustering.

A wide assortment of closeness and disparity measures exists. The measures, for example, cosine, Jaccard coefficient, Pearson Correlation Coefficient are similitude measures where as the separation measures like Euclidian, Manhattan, Minkowski are difference measures.

These measures have been proposed and broadly connected for record clustering.

Similarity measure can be converted into dissimilarity measure:

Dissimilarity=1-Similarity (4)

Cosine Similarity

The similarity between the two documents d$_i$, d$_j$ can be calculated using cosine as

$$\cos(d_i, d_j) = \frac{\sum_{k=1}^{n}(d_{i,k} * d_{j,k})}{\sqrt{\sum_{k=1}^{n}(d_{i,k})^2 * \sum_{k=1}^{n}(d_{j,k})^2}} \quad (5)$$

Where n represent the number of terms. When the cosine value is 1 the two documents are identical, and 0 if there is nothing in common between them (i.e., their document vectors are orthogonal to each other).

Jaccard Similarity

The Cosine Similarity may be extended to yield Jaccard Coefficient

$$jaccard(d_i, d_j) = \frac{\sum_{k=1}^{n}(d_{i,k} * d_{j,k})}{\sum_{k=1}^{n} d_{i,k} + \sum_{k=1}^{n} d_{j,k} - \sqrt{\sum_{k=1}^{n}(d_{i,k} * d_{i,k})}} \quad (6)$$

Euclidean Distance

This Euclidean distance between the two documents d$_i$, d$_j$ can be calculated as

$$Euclidean\ Distance(d_i, d_j) = \sqrt{\sum_{k=1}^{n}(d_{ik} - d_{jk})^2} \quad (7)$$

Where, n is the number of terms present in the vector space model.

Euclidian distance gives the dissimilarity between the two documents. If the distance is smaller it indicates they are more similar else dissimilar.

Manhattan Distance

The Manhattan distance between the two documents d$_i$, d$_j$ can be calculated as

$$\text{Manhattan Distance}(d_i,d_j)=\sum_{k=1}^{n}|d_{ik}-d_{jk}| \qquad (8)$$

Links and Neighbors :

Two reports are thought to be neighbors on the off chance that they are like each other [6] and the connection between the archives speak to the quantity of their normal neighbors.

Let sim(docu,docv) figures match shrewd archive similitude and ranges in [0, 1], esteem one shows docu, docv are similar and zero demonstrates that records docu, docv are distinctive. On the off chance that Sim(docu, docv) ≥ with incentive in the vicinity of 0 and 1 then docu, docv are neighbors, where is determined by the client to demonstrate the comparability among archives to be neighbors. At the point when is set to 1, then it is a neighbor of another precisely same report and if is set to zero then any record can be its neighbor.

Consequently esteem ought to be set precisely. In this work subsequent to performing many examinations with various datasets we have touched base to a conclusion to set programmed an incentive for . For a picked similitude esteem x, include of sections ≥x likeness grid is 2 times N where N is the measure of dataset then closeness esteem for can be set as x.

Neighbor of each record is spoken to in a framework called as neighbor lattice. Give NM a chance to be a n x n framework of neighbors with n being the dataset measure and in light of docu, docv being neighbors NM[u,v] is set to 1 or 0 [7]. Let N[docu] gives the check of neighbors of docu got from NM with uth push sections as one.

The links(docu, docv) is utilized to discover the check of shared neighbors of docu, docv[6] and is computed as a result of uth line, vth section of NM.

$$\text{n}\, link(doc_u,doc_v)=\sum_{m=1}^{n} NM[u,m]\times NM[m,v].......5$$

Thus, large value of link(doc$_u$, doc$_v$) has high possibility of these documents assigned to one cluster. Since the measures [Cosine/Jaccard/Pearson] measure pair wise similarity between two documents, these measures alone will lead to general or narrow clustering while using link function with these measures can be considered as a specific or comprehensive clustering approach [6], as neighbor data in similarity adds global view to determine documents similarity.

NEIGHBORS ALGORITHM

K-implies calculation is utilized to group reports into k number of parcels. In K-implies calculation, at first k-items are chosen haphazardly as centroids. At that point relegate all items to the closest centroid to shape k-groups. Process the centroids for each group and reassign the articles to shape k-bunches by utilizing new centroids. Processing the centorids and reassigning the articles ought to be rehashed until there is no adjustment in the centroids. As the underlying k-articles are chosen arbitrarily relying upon the choice of these k-questions the exactness and effectiveness of the classifier will differ. Rather than choosing the underlying centroids haphazardly, we are proposing to locate the best beginning centroids. The fundamental goal for picking the best beginning centroids is to diminish the quantity of emphasess for the parceling based calculations in light of the fact that the quantity of cycles to get the last bunches relies on upon the underlying centroids picked. On the off

Volume No:2, Issue No:11 (April-2017) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

chance that the quantity of emphasess diminished then the productivity of the calculation will be expanded. In the proposed strategy the underlying centroids are picked by utilizing most remote neighbors. For the trial reason the calculation picked is k-implies which is one of the outstanding dividing based clustering calculations. To expand the proficiency of the k-implies calculation as opposed to choosing k-questions haphazardly as starting centroids the k-items are picked by utilizing most remote neighbors. Subsequent to finding the underlying centroids by utilizing most distant neighbors apply k-implies calculation to bunch the reports.

The reports should be preprocessed before applying the calculation. Evacuating of stop words, performing stemming, pruning the words that show up with low recurrence and so on., are the preprocessing steps. In the wake of preprocessing vector space model is assembled.

The calculation works with disparity measures. The archives are more comparative if the separation between the records is less else the reports are unique. Calculation for finding the underlying centroids by utilizing most remote neighbors is as per the following:

Algorithm:

1.By utilizing the difference measures build disparity framework for the archive combines in the vector space show.

2.Find the greatest incentive from the uniqueness grid

3.Find the report match with the greatest esteem found in step 2 and pick them as initial two starting (i.e., these two archives are the most remote neighblrs)

4.For finding staying indicated number of introductory centroids

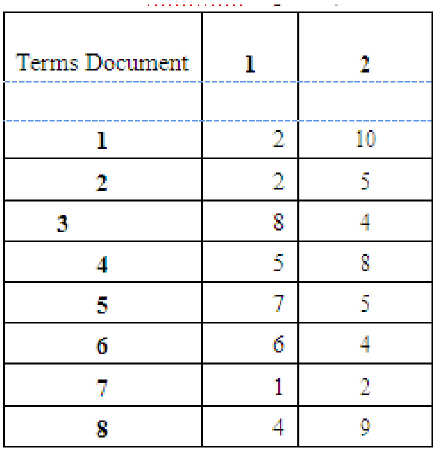
- Calculate the centroid for officially discovered beginning centroids.
- With the centroid figured in step 4.i produce the disparity network amongst centroid and every one of the records aside from those picked as introductory centroids.
- Find the most extreme incentive from the divergence lattice created in step 4.ii and pick the comparing record as next starting centroid.

5.Repeat the progression 4 until the predefined number of starting centroids are picked.

Frame the vector space demonstrate given beneath build 3 bunches by utilizing k-implies calculation and utilize the most remote neighbors for finding the underlying centroids.

The calculation is clarified as takes after for finding the underlying centroids:

Step1: Consider term recurrence vector space display for 8 reports and create disparity lattice utilizing uniqueness measures. (For the clarification we have considered manhattan separate)



Terms Document	1	2
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

TABLE 1: Vector Space Model.

Volume No:2, Issue No:11 (April-2017) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Documents	1	2	3	4	5	6	7	8
1	0	5	12	5	10	10	9	3
2	5	0	7	6	5	5	4	6
3	12	7	0	7	2	2	9	9
4	5	6	7	0	5	5	10	2
5	10	5	2	5	0	2	9	7
6	10	5	2	5	2	0	7	7
7	9	4	9	10	9	7	0	10
8	3	6	9	2	7	7	7	0

TABLE 2: Dissimilarity Matrix.

Step 2: The maximum value from the dissimilarity matrix is 12.

Step 3: The document pair with the maximum value is (1,3) and there for the first two initial centroids are (2,10) and (8,4) which represents document 1 and 3 respectively.

Step 4: As mentioned in the problem, 3 centroids are required. Already 2 centroids are choosen from step 3. The remaiming 1 centorid need to be found.

Step 4.i: the centroid for (2,10) and (8,4) is (5,7)

Step 4.ii: Generating the dissimilarity matrix between (5,7) and (2,5), (5,8), (7,5), (6,4), (1,2), (4,9) which represents the documents 2,4,5,6,7 and 8 respectively

Step 4.iii: maximum value is 9 and the corresponding document is 7. Therefore the third initial centroid is (1,2)

Step 5: Required number of initial centroids is found so no need of repeating step 4.

Now by applying the k-means algorithm by using the initial centroids obtained above the clusters are formed are shown in fig 5.

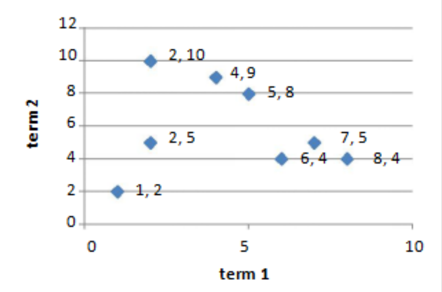


FIGURE 1: D documents plotting

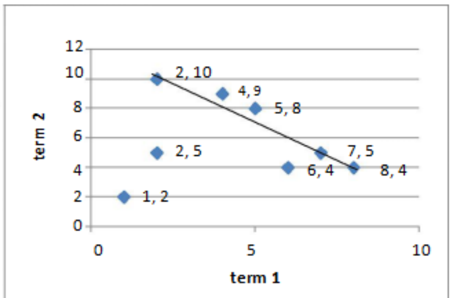


FIGURE 2: Finding the farthest neighbors

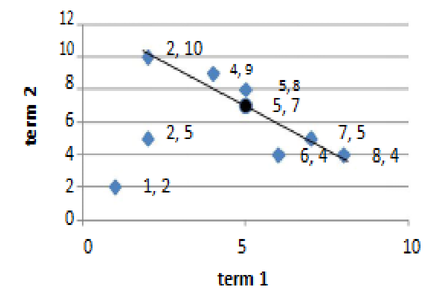


FIGURE 3: find the centroid of the farthest neighbors

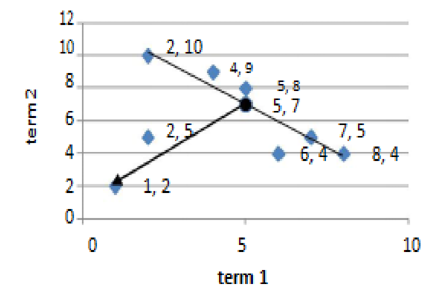


FIGURE 4: find the farthest point from the centroid

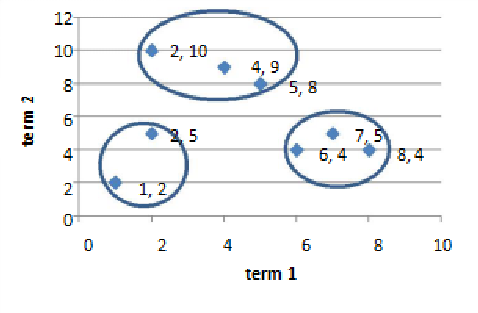
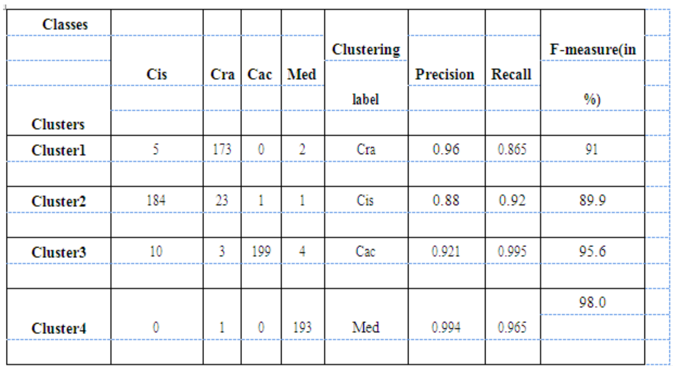


Figure 5: the final three clusters after applying k- means

Pre-Processing

Before building the datasets for our examinations we disposed of archives with single word record measure. For the datasets considered figured normal record estimate and disregarded those archives that are not as much as normal document measure. On every class we have connected the record decrease technique where we considered the archives of a classification in the dataset fulfilling normal document measure and dispensed with different reports of the classification. To accomplish this we manufactured a Boolean vector space portrayal of records where for every classification normal document size is resolved and pruned archives that are with length not as much as normal record measure, in this manner shaping substantial reports. On these reports we connected preprocessing which incorporates, tokenization of info record, expulsion of uncommon characters, evacuation of stop words, connected stemming to infer stem words, recognized exceptional terms and fabricated a vector of term archive portrayal. At that point we ascertained archive recurrence of all terms and expelled less incessant terms from the vector as the incorporation of these terms shape groups of little sizes. The terms with high recurrence of event are additionally pruned for they won't add to clustering process.

RESULTS



Classes Clusters	Cis	Cra	Cac	Med	Clustering label	Precision	Recall	F-measure(in %)
Cluster1	5	173	0	2	Cra	0.96	0.865	91
Cluster2	184	23	1	1	Cis	0.88	0.92	89.9
Cluster3	10	3	199	4	Cac	0.921	0.995	95.6
Cluster4	0	1	0	193	Med	0.994	0.965	98.0

TABLE 3: Clustering the documents by using k-means, jaccard similarity measure and random selection ofinitial centroids.

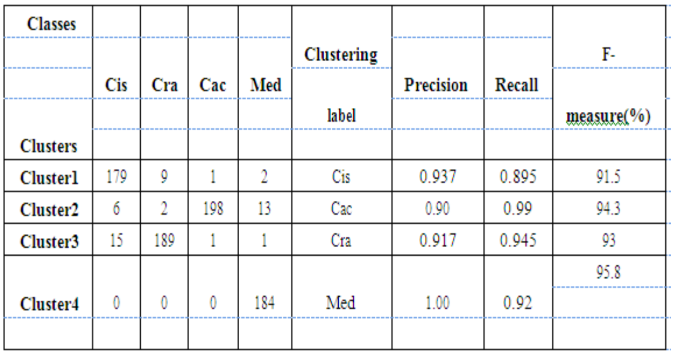
The overall accuracy=average of all F-measure

i.e., **accuracy=93.60**

One iteration, includes computing the centroids and assigning the objects to the predefined number of clusters.

These iterations are repeated until consecutive iterations yield same centroids.

Iterations =18



Classes Clusters	Cis	Cra	Cac	Med	Clustering label	Precision	Recall	F-measure(%)
Cluster1	179	9	1	2	Cis	0.937	0.895	91.5
Cluster2	6	2	198	13	Cac	0.90	0.99	94.3
Cluster3	15	189	1	1	Cra	0.917	0.945	93
Cluster4	0	0	0	184	Med	1.00	0.92	95.8

TABLE 4: Clustering the documents by using k-means, jaccard similarity measure and farthest neighborsas initial centroids.

Accuracy=93.65

Iterations= 5

Classes								
	Cis	Cra	Cac	med	Clustering label	Precision	Recall	F-measure
Clusters								
Cluster1	1	137	1	61	Cra	0.685	0.685	68
Cluster2	0	0	194	1	Cac	0.994	0.97	98
Cluster3	197	63	5	4	Cis	0.732	0.985	83.9
Cluster4	1	0	0	134	Med	0.99	0.67	79.9

TABLE 5: Clustering the documents by using k-means, Cosine similarity measure and random selection ofinitial centroids.

Accuracy=82.45

Iterations= 44

Classes								
	Cis	Cra	Cac	Med	Clustering label	Precision	Recall	F-measure(%)
Clusters								
Cluster1	1	152	0	31	Cra	0.826	0.76	79.1
Cluster2	0	0	188	2	Cac	0.989	0.94	96.4
Cluster3	199	48	12	4	Cis	0.75	0.995	85.5
Cluster4	0	0	0	162	Med	1.00	0.81	89.5

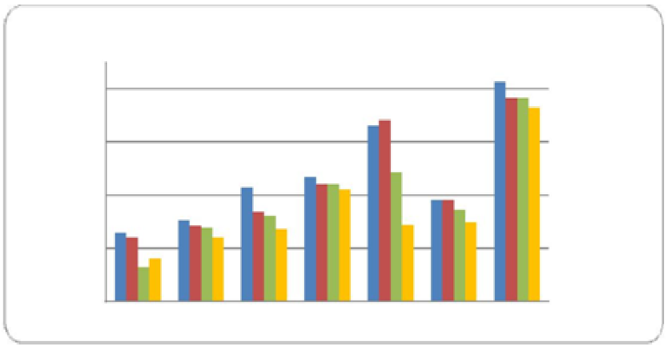
TABLE 6: Clustering the documents by using k-means,cosine similarity measure and farthest neighbours asinitial centroids.

Accuracy=87.62

Iterations= 7

Firstly, performance of initial centers is considered, following it, automatic determination of number of partitions in a given dataset, are described, next we see the performance of SBTC, RBTC and SNBTC approaches, where SBTC is Simple Bisecting K-means Text Clustering, RBTC is Rank Based Text [9] implemented for kmeans is extended to bisecting k-means in this work and proposed SNBTC, Shared Neighbor Based Text Clustering are analyzed and lastly we compare the effect of applying proposed approaches in clustering algorithm.

Figure 1 depicts the performance of Initial centers methods, where in Sequential, Random, Rank Neighbors and Shared Neighbor based are compared. For each type of initial centers choosen, we have run Bisecting K-means clustering, and the quality of the clusters formed are evaluated. To compare the results we used entropy as the quality measure. The lesser the value of entropy, the better is its quality, and the proposed shared neighbor seed document selection method, has showed significant improvement in the clustering process.



Results of Bisecting K-means with Initial Centers

Different k values, where thematic cohesive clusters are expected to form. At these k-values simple bisecting k-means is applied and observed intra cluster similarity to be maximum at these k's. The experiments showed quite accurate results.

CONCLUSIONS

In this paper an attempt is made to improve performance of bisecting k-means. This work has given a neighbor based solution to find number of partitions in a dataset.

Then, proceeds to give an approach to find k initial centres for a given dataset. The family of k-means require k initial centers and number of clusters to be specified. In this work we have addressed these two issues with neighbor information. Then we proposed a heuristic measure to find the compactness of a cluster and when employed in selecting the cluster to be split in bisecting step has shown improved performance. All the three approaches proposed, when applied to bisecting k-means shown better performance.

We have experimented with neighbors and links concept specified in and found that the cluster quality improves with neighbor information combined with text clustering similarity measures. Neighbors are used in determining the compactness of clusters in bisecting k-means. In our previous study we have noticed that Jaccard and Cosine outperforms Pearson coefficient with link function. It is observed that the clusters formed are cohesive. Efficiency of clustering results are based on representation of documents, measure of similarity and clustering technique. In our future work semantics knowledge shall be incorporated in the document representation to establish relations between tokens and study various measures semantic and similarity on these representations with neighbors based clustering approaches for better clustering results.

Here we proposed new method for finding initial centroids by using farthest neighbors. The experimental results showed that accuracy and efficiency of the k-means algorithm is improved when the initial centroids are chosen using farthest neighbors than random selection of initial centroids. As the number of iterations decreased we can tell the efficiency is improved. We intend to apply this algorithm for different similarity measures and study the effect of this algorithm with different benchmark datasets exhaustively. In our future work we In this paper an endeavor is made to enhance execution of bisecting k-implies. This work has given a neighbor based answer for discover number of parcels in a dataset. At that point, continues to give a way to deal with discover k beginning places for a given dataset. The group of k-means require k starting focuses and number of bunches to be determined. In this work we have tended to these two issues with neighbor data. At that point we proposed a heuristic measure to discover the conservativeness of a bunch and when utilized in choosing the group to be part in bisecting step has demonstrated enhanced execution. All the three methodologies proposed, when connected to bisecting k-implies indicated better execution. We have explored different avenues regarding neighbors and connections idea determined in and found that the bunch quality enhances with neighbor data consolidated with content clustering comparability measures. Neighbors are utilized as a part of deciding the smallness of groups in bisecting k-implies. In our past review we have seen that Jaccard and Cosine beats Pearson coefficient with connection work. It is watched that the groups framed are durable.

Proficiency of clustering results depend on portrayal of archives, measure of similitude and clustering system. In our future work semantics learning should be fused in the archive portrayal to set up relations amongst tokens and study different measures semantic and likeness on these portrayals with neighbors based clustering approaches for better clustering outcomes.

Here we proposed new technique for discovering starting centroids by utilizing most remote neighbors. The test comes about demonstrated that precision and proficiency of the k-implies calculation is enhanced when the underlying centroids are picked utilizing most distant neighbors than arbitrary choice of beginning centroids. As the quantity of emphasess diminished we can tell the productivity is progressed. We mean to apply this calculation for various likeness measures and study the impact of this calculation with various benchmark datasets thoroughly. In our future work we additionally expect to investigate alternate procedures for picking the best beginning centroids and apply them to partitional and hierarchal clustering calculations to enhance the effectiveness and precision of the clustering calculations.

REFERENCES

[1] O. Zamir, O. Etzioni, O. Madani, R.M. Karp, "Fast and intuitive clustering of web documents", in: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997, pp. 287–290.

[2] C.C. Aggarwal, S.G. Gates, P.S. Yu, "On the merits of building categorization systems by supervised clustering", in: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp.352–356.

[3] B. Larson, C. Aone, "Fast and effective text mining using linear-time document clustering", in: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 98(463), 1999, pp. 16–22.

[4] Salton, G., Wong, A., Yang, C.S. (1975). "A vector space model for automatic indexing". Communications of the ACM, 18(11):613-620.

[5] na Huang, "Similarity Measures for Text Document Clustering", published in the proceedings of New Zealand Computer Science Research Student Conference 2008.

[6] Saurabh Sharma, Vishal Gupta. "Domain Based Punjabi Text Document Clustering". Proceedings of COLING 2012: Demonstration Papers, pages 393–400,COLING 2012,Mumbai, December 2012.

[7] D. Manning, Prabhakar Raghavan, Hinrich Schütze, "An Introduction to Information RetrievalChristopher", Cambridge University Press, Cambridge, England

[8] M.F. Porter,Analgorithm" for suffix stripping",Program, vol.14, no.3, pp. 130−137, 1980.

[9] C.J.Van Rijsbergen,(1989), "Information Retrieval", Buttersworth, London, Second Edition.

[10] G. Kowalski,"Information Retrieval Systems–Theory and Implementation", Kluwer Academic Publishers, 1997.