ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

A Relevant Document Information Clustering Algorithm for Web Search Engine

Y.SureshBabu, K.Venkat Mutyalu, Y.A.Siva Prasad

Abstract— **Search engines are the Hub of Information, The advances in computing and information storage have provided vast amount of Data, the users of World Wide Web is increasingly day by day, It is become more difficult to users get the required information according to their interests. The IR community has explored document clustering as an alternative method of organizing retrieval results, so by using clustering concept we can find the grouped relevant documents. The purpose of clustering is to partitioning the set of entities into different groups called clusters. These groups may consistent in terms of similarity of its members. As the name suggests, the representative based clustering techniques uses some form of representation for each cluster. Thus every group has a member that represents it. The main use is to increase the efficiency of the algorithm and to decrease the cost of the algorithm. Clustering process is done by using k-means partitioning algorithms and Hierarchical clustering algorithms but there are lot of disadvantages, it works very slow and it is not applicable for large databases. So fast greedy k-means algorithm is used it overcomes the drawbacks of k-means algorithm and it is very much accurate and efficient. So we introduce an efficient method to calculate the distortion for this algorithm. This helps the users to find the relevant documents more easily than by relevance ranking.**

Index Terms—**Information retrieval, K-Means, Fast k-means, Document clustering, Web clustering.**

I.INTRODUCTION

Search engines are the Hub of Information, The advances in computing and information storage have provided vast amount of Data, the users of World Wide Web is increasingly day by day, It is become more difficult to users get the required information according to their interests .In recent years, outsized archives of data are available in industry and organization, which results from the accumulation of data. Utilizing these bulk data for decision-making may lead to decisive problems. Finding the right information from such a large collection is extremely difficult.

A user's interaction with the search output is often far from optimal. If the few sampled documents are not found relevant, it is very likely that the rest of the documents will not be inspected at all. Clustering the search output could also prove useful in automatic query expansion. If clustering could achieve a helpful categorisation of the documents, users could be able to base query expansion on certain clusters that seem more likely to lead to aspects so far unrepresented among the retrieved documents.

Clustering is an unsupervised classification technique. A set of unlabeled objects are grouped into meaningful clusters, such that the groups formed are homogeneous and neatly separated. Challenges for clustering categorical data are: 1) Lack of ordering of the domains of the individual attributes. 2) Scalability to high dimensional data in terms of effectiveness and efficiency. One of the techniques that can play an important role towards the achievement of this objective is document clustering. The increasing importance of document clustering and the variety of its applications has led to the development of a wide range of algorithms with different quality. Based on this model, the following are the key requirements for Web document clustering methods.

- **Relevance**: The method ought to produce clusters that group documents relevant to the user's query separately from irrelevant ones.
- **Browsable Summaries**: The user needs to determine at a glance whether a cluster's contents are of interest. Sifting through ranked lists is not replaced with sifting through clusters. Therefore the method has to provide concise and accurate descriptions of the clusters.
- **Overlap**: Since documents have multiple topics, it is important to avoid confining each document to only one cluster.
- **Speed**: The clustering method ought to be able to cluster up to one thousand snippets in a few seconds. For the impatient user, each second counts.

Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians, in which the objective is to minimize the sum of distances to the nearest center, and the geometric k-center problem, in which the objective is to minimize the maximum distance from every point to its closest center. The

.
Y.Suresh Babu*, Pursuing M.Tech(cse),Sri Vasavi Engineering College,(Affiliated to JNTU Kakinada). Tadepalligudem, AndhraPradesh.*

Mr.K.Venkat Mutyalu, *Associate Professor, Dept of IT, Sri Vasavi Engineering College,(Affiliated to JNTU Kakinada,.Tadepalligudem, Andhra Pradesh.*

Mr.Y.A.Siva Prasad, *Research Scholar in CSE Dept, KL University, Andhra Pradesh.*

fast greedy k-means algorithm overcomes the major shortcomings of the k-means algorithm discussed above, possible convergence to local minima and large time complexity with respect to the number of points in the dataset. One group of experiments aimed to assess the ability of the implementation to bring together topically related documents Implementation included a procedure of term selection for document representation which preceded the clustering process and a procedure involving cluster representation for users' viewing following the clustering process. After some tuning of the implementation parameters for the databases used, several different types of experiments were designed and conducted to assess whether clusters could group documents in useful ways.

A. Offline Clustering

The application that demonstrates the basic offline clustering task. Provides k-means and bisecting k-means partitioned clustering. It will run each algorithm on the first 100 documents in the index (or all of them if less than 100) and print out the results. The parameters accepted by Offline Cluster are:

- Index -- the index to use. Default is none.
- Cluster Type -- Type of cluster to use, either agglomerative or centroid. Centroid is agglomerative using mean which trades memory use for speed of clustering. Default is centroid.
- SimType -- The similarity metric to use. Default is cosine similarity (COS), which is the only implemented method.
- DocMode -- The integer encoding of the scoring method to use for the agglomerative cluster type. The default is max (maximum). The choices are:
- max -- Maximum score over documents in a cluster.
- mean -- Mean score over documents in a cluster. This is identical to the centroid cluster type.
- avg -- Average score over documents in a cluster.
- min -- Minimum score over documents in a cluster.
- NumParts -- Number of partitions to split into. Default is 2
- MaxIters -- Maximum number of iterations for k-means. Default is 100.
- BkIters -- Number of k-means iterations for bisecting k-means. Default is 5.

B. Online Clustering

In this by using internet and by taking open source search engine by using java package lucene and the elements taking in a 2dimentional array (link, key element position) in a document by using Euclidean space model we find out the distance ie..correlation of the document and made it as a group if similar.

II.DOCUMENT CLUSTERING

Document clustering analysis plays an important role in document mining research. Document clustering has been traditionally investigated mainly as a means of improving the performance of search engines by pre-clustering the entire corpus (the cluster hypothesis - van Rijsbergen, 79).Clustering is one of the main data analysis techniques and deals with the organization of a set of objects in a multidimensional space into unified groups, called clusters. Each cluster contains objects that are very similar to each other and very dissimilar to objects in other clusters. Cluster analysis aims at discovering objects that have some representative behaviour in the collection. Clustering is a form of unsupervised classification, which means that the categories into which the collection must be partitioned are not known. In order to cluster documents, one must first choose the type of the characteristics or attributes (e.g. words, phrases or links) of the documents on which the clustering algorithm will be based and their representation. The most commonly used model is the Vector Space Model. Vector Space Model is a mathematical model to represent Information Retrieval Systems which uses term sets to represent both documents and queries, employs basic linear algebra operations to calculate global similarities between them.

Numerous documents clustering algorithms appear in the literature. A widely adopted definition of optimal clustering is a partitioning that minimizes distances within a cluster and maximizes distances between clusters. In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically from the documents to be clustered, and the documents are subsequently assigned to those clusters. Users are known to have difficulties in dealing with information retrieval search outputs especially if the outputs are above a certain size. It has been argued by several researchers that search output clustering can help users in their interaction with IR systems. The utility of this data set was limited for various reasons however, it can be concluded that clusters cannot be relied on to bring together relevant documents assigned to a certain facet. While there was some correlation between the cluster and facet assignments of the documents when the clustering was done only on relevant documents, no correlation could be found when the clustering was based on results of queries defined by City participants to the Interactive track.

A. Document representation

Vector space model is the most commonly used document representation model in text and web mining area. In this model, each document is represented as an n-dimensional vector. The value of each element in the vector reflects the importance of the corresponding feature in the document. As mentioned above, document features are unique terms. After the above transformation, the complicated, hard-to-understand documents are converted into machine acceptable, mathematical representations. The problem of measuring the similarity between documents is now converted to the problem of calculating the distance between document vectors. The Document frequency (DF) of a term is the number of documents in which that term occurs. One can use DF as a criterion for selecting good terms. The basic intuition behind using document frequency as a criterion is that rare terms either do not capture much information about one category, or they do not affect global performance. In spite of its simplicity, it is believed to be as effective as more advanced feature selection methods. According to Korpimies&Ukkonen term weighting is necessary in output clustering and the focus should be on the term frequencies within the output set; terms which are frequent or too infrequent within the document set should be given small weights. They have formulated "Contextual inverted document frequency" as:

$$cidf(Q,t_j) = \frac{1}{\sum_{i=1}^{m} w_{ij} x rel(Q,D_i)} \quad (1)$$

B. Measuring the association between documents

The most common measures of association used in search engine are:
1. Simple matching coefficient: number of shared index terms,
2. Dice's coefficient: the number of shared index terms divided by the sum of the number of terms in two documents. If subtracted from 1, it gives a normalized symmetric difference of two objects.
3. Jaccard's coefficient: number of shared index terms divided by union of terms in two documents,
4. Cosine coefficient: number of shared index terms divided by multiplication of square roots of number of terms in each document,
5. Overlap coefficient: number of shared index terms divided by minimum of number of terms in each document.

There are also several dissimilarity coefficients, Euclidian distance being the best known among them. However, it has a number of important shortcomings: It is scale dependent, which may cause serious problems when it is used with raw data and it assumes that the variable values are uncorrelated with each other. A major limitation in the IR context is that it can lead to two documents being regarded as highly similar to each other, despite the fact that they share no terms in common(but have lots of negative matches). The Euclidian distance is thus not widely used for document clustering, except in Ward's method.

III. CHOICE OF METHOD FOR DOCUMENT CLUSTERING

A. K-means Algorithm

K-Means clustering is a very popular algorithm to find the clustering in dataset by iterative computations. It has the advantages of simple implementing and finding at least local Optimal clustering. K-Means algorithm is employed to find the clustering in dataset. The algorithm is composed of the following steps:

Method:
(1) Arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar,based on the mean value of the objects in the cluster;
(4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

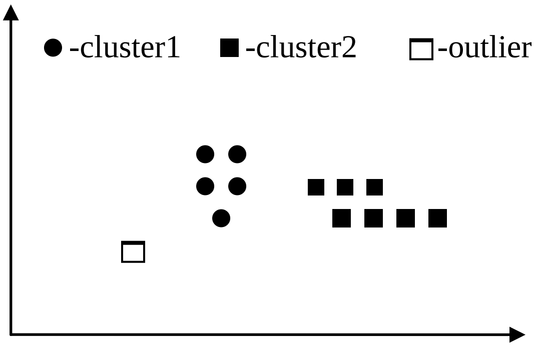


Fig1:Samples in 2D

Here *x*1 = (2,3), *x*2 = (4,9), *x*3 = (8,15)
Where *x*4 = (12,7), *x*5 = (13,10)
Because we want to produce 2 clusters of these samples, set *k*=2 Our steps are:
1. Set initial points. Because *k*=2 , we select two points *c*1=*x*1 and *c*2=*x*2 as center points.
2. *x*2 is near *c*1, so put in to cluster1. Now new 2 center are (3,6) and (11,10.67) for each sample ,find its nearest center (don't recomputed the centers .sample *x*1 and *x*2 are near (3,6) sample *x*3,*x*4 and *x*5 are near (11,10.67). Members of each cluster don't change . so stop

The k-means clustering algorithm is one of the popular data clustering approaches. The kmeans clustering algorithm receives as input a set of (in our case 2- dimensional) points and the number *k* of desired centers or cluster representatives. With this input, the algorithm then gives as output a set of point sets such that each set of points have a defined center that they “belong to” that minimizes the distance to a center for all the possible choices of each set.

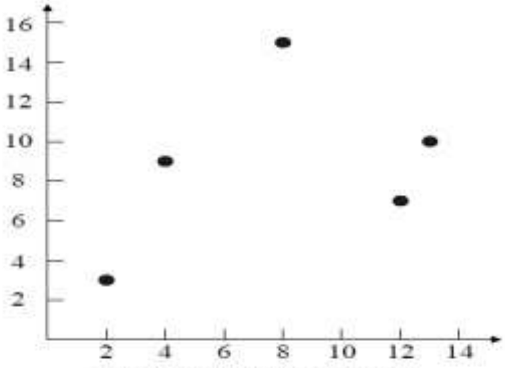


Fig.2 An outlier of samples

B. The Fast Greedy k-means Algorithm

With K-Means algorithm, different initial cluster center can lead to different times of iterative operations, which brings about different algorithm efficiency. The fast greedy k-means algorithm is similar to Lloyd’s in that is searches for each point the best center of gravity to “belong to” but different in the assignments. Lloyd’s algorithm, in each iteration, reassigns up to every point to a new center and then readjusts the centers accordingly then repeats. The Progressive Greedy approach does not act upon every point in each iteration, rather the point which would most benefit moving to another cluster. The following 4 steps that have been illustrated outline the algorithm.

(1)Construct an appropriate set of positions/locations which can act as good candidates for insertion of new clusters;

(2)Initialize the first cluster as the mean of all the points in the dataset;

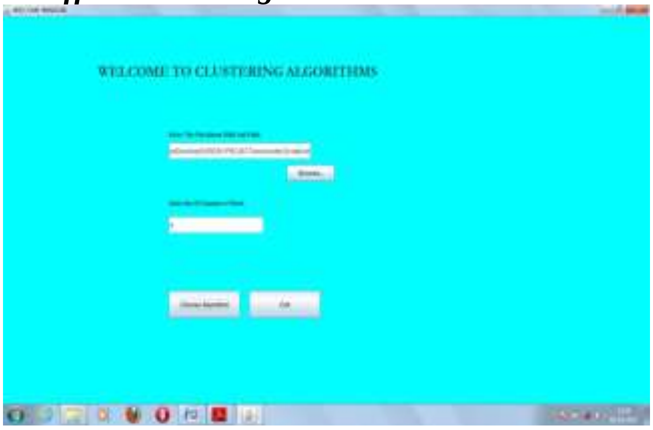
(3)In the *k*th iteration, assuming K-1 clusters after convergence find an appropriate position for insertion of anew cluster from the set of points created in step 1 that gives minimum distortion,

(4)Run k-means with K clusters till convergence. Go back to step 3 if the required number of clusters is not yet reached .

IV.RELATED WORK

First we have to give any keyword in the search engine, then we will get results related to that keyword. For those results we will calculate the similarities of the documents and put it into one cluster. Similarities can be measured by using k means algorithm which uses euclidiean space model distance measure. For example I am taking here 2dimentioanl data (x,y) as document (link,position of the keyword). I am taking 2-dimentional data as (1,3),(1,6),(1,7),(2,8). here 1 means 1st link in that 3rd position of the keyword. There by calculating similarities and if the results are similar to each other that means the distance between the two documents is minimum then we will put those documents into one cluster. So it is a intra clustering. Minimizing the distance between the data points in the cluster and maximizing the distance between the clusters.

A. Offline clustering:



WELCOME TO CLUSTERING ALGORITHMS

Fig 3:Shows the clustered input by K-Means

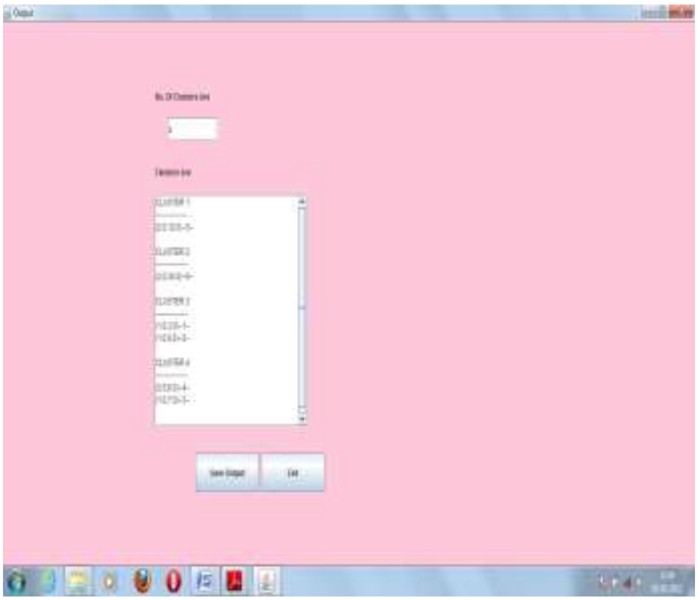


Fig:4 Shows the clustered output by k-means algorithm.

B.ONLINE CLUSTERING

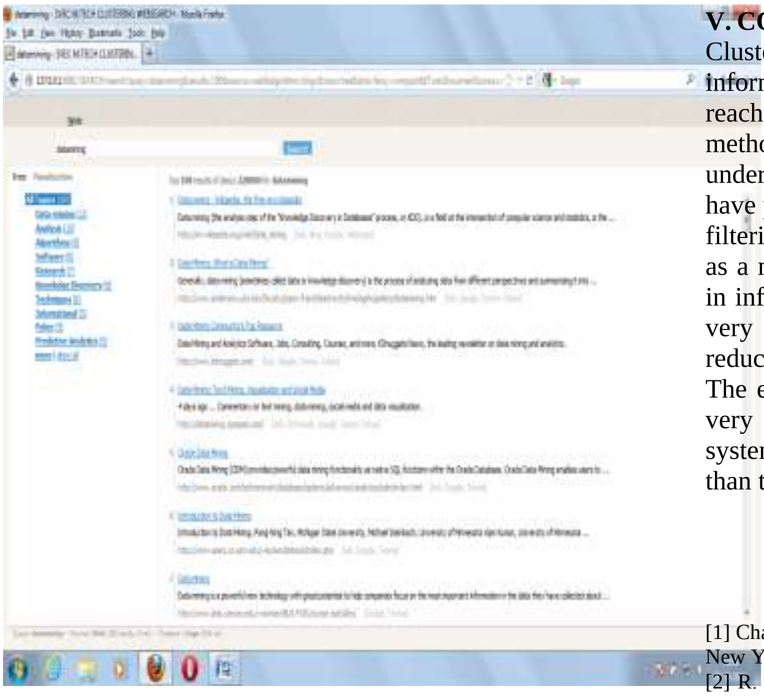


Fig5: Tree view Results obtained for the given keyword ‘DataMining’

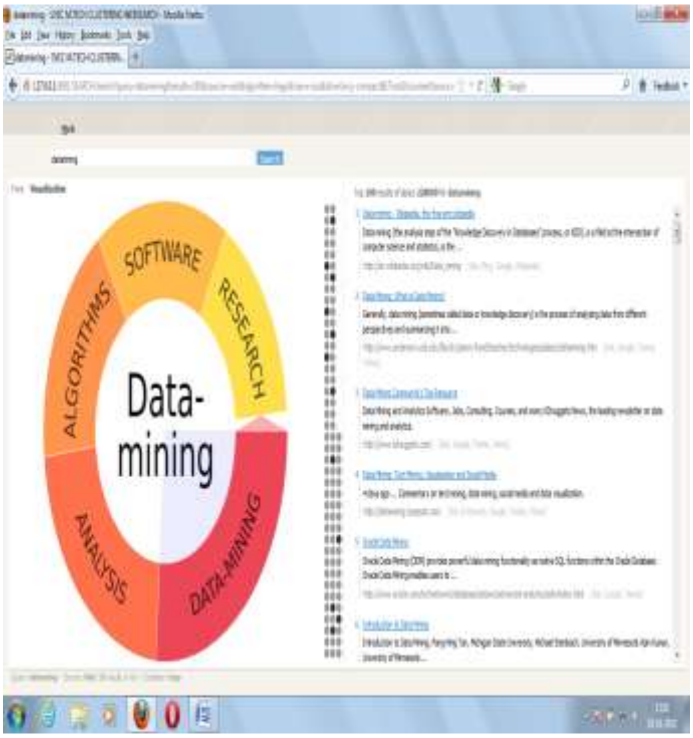


Fig: 6 Data Visualization showing the related clustered group of results.

V. CONCLUSION

Clustering can increase the efficiency and the effectiveness of information retrieval. Clustering is an efficient way of reaching information from raw data and K-means is a basic method for it. Although it is easy to implement and understand, K-means has serious drawbacks. In this paper we have presented an efficient method of combining the restricted filtering algorithm and the greedy global algorithm and use it as a means of improving user interaction with search outputs in information retrieval systems. Thus document clustering is very useful to retrieve information application in order to reduce the consuming time and get high precision and recall The experimental results suggest that the algorithm performs very well for Document clustering in web search engine system and can get better results for some practical programs than the ranked lists and k-means algorithm.

REFERENCES

[1] Chan, L.M.: Cataloging and Classification : an Introduction. McGraw-Hill, New York, 1994

[2] R. Kannan, S. Vempala, and Adrian Vetta, “On Clusterings: Good, Bad, and Spectral”, Proc. of the 41st Foundations of Computer Science, Redondo Beach, 2000.5

[3] S. Kantabutra, Efficient Representation of Cluster Structure in Large Data Sets, Ph.D. Thesis, Tufts University, Medford,MA, September2001.

[4] Dan Pelleg and Andrew Moore: X-means: Extending kmeans with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, Palo Alto, CA, July 2000.

[5] Aristides Likas, Nikos Vlassis and Jacob J. Verbeek: The global k-means clustering algorithm. In Pattern Recognition Vol 36, No 2, 2003.

[6] J. Matoušek. On the approximate geometric k-clustering. Discrete and Computational Geometry. 24:61-84, 2000

[7] Dan Pelleg and Andrew Moore: Cached sufficient statistics for efficientmachine learning with large datasets. In Journal of Artificial Intelligence Research, 8:67-91, 1998.

[8]A Document Clustering Algorithm for Web Search Engine Retrieval System ,2010 Hongwei Yang School of Software,Yunnan University, Kunming 650021, China; Education Science Research Academy of Yunnan, Kunming 650223,China.

[9] C. J. van Rijsbergen, Information Retrieval, Butterworths, London, 2nd ed., 1979.

First Author: Mr.Y.SURESH BABU pursuing his M.Tech (CSE) in Sri Vasavi Engineering College (Affiliated to JNTU Kakinada), Tadepalligudem, Andhra Pradesh.

Second Author: Mr.K.Venkat Mutyalu working as Associate Professor in Dept of IT,Sri Vasavi Engineering College(Affiliated to JNTU Kakinada), Tadepalli Gudem, AndhraPradesh.

Third Author : Mr.Y.A Siva Prasad, Research Scholar in CSE Department,, KL University, Andhra Pradesh, and Life member of **CSI, IAENG**, Having 9 years of Teaching Experience and presently working as an Associate Professor at Chendhuran College of Engineering & Technology ,Pudukkotai, Tamilnadu .