Document Clustering in Web Search Engine

A.S.N.Chakravarthy[*1], Deepthi.S[#2], K.Satyatej[#3], Sk.Nizmi[#4], S.Sindhura[#5]

[*] *Professor, Department of Electronics and Computer engineering,*
K.L.University,Guntur,A.P.,India.
[#] *Student (B.Tech), Department of Electronics and Computer engineering,*
K.L.University,Guntur,A.P.,India.

Abstract— As the number of web pages grows, it becomes more difficult to find the relavant documents from the information retrieval engines, so by using clustering concept we can find the grouped relavant documents. The main purpose of clustering techniques is to partitionate a set of entities into different groups, called clusters. These groups may be 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 reduce the cost of the algorithm, the use of representatives makes the process easier to understand. The most popular Clustering technique is the k-means algorithm where it has a lot of disadvantages, it works very slow and it is not applicable for large databases. So fast greedy k-means algorithm is used, which overcomes the drawbacks of k-means algorithm and it is very much accurate and efficient. So we introduce an efficient method to compute the distortion for this algorithm.

Key Terms —Document clustering, k-means, Fast k-means algorithm.

I .INTRODUCTION

A. Offline Clustering

The application that demonstrates the basic offline clustering task. Provides k-means and bisecting k-means partitional 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.
- sim Type -- 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. 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 [6]. They have formulated "Contextual inverted document frequency" as:

$$cidf(q,t) = \frac{1}{\sum_{n-1}^{n} w_{ij \times rel(Q,D_i)}} \quad (1)$$

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 is 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

Following steps:

(1) Initialize k cluster centers to be seed points (These Centers can be randomly produced or use other Ways to generate).
(2) For each sample, find the nearest cluster center, put the sample in this cluster and recompute centers of the altered cluster (Repeat n times).
(3) Examine all the samples again and put each one in the cluster identified with the nearest center (don't recompute any cluster centers).
(4) If members of each cluster haven't been changed, stop. If changed, go to step 2.

- K-Means clustering is a very popular algorithm to find the clustering in dataset by iterative computations,Partitional clustering approach.
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid.

Given an example of k-means algorithm using data shown on the following (see Fig.1):

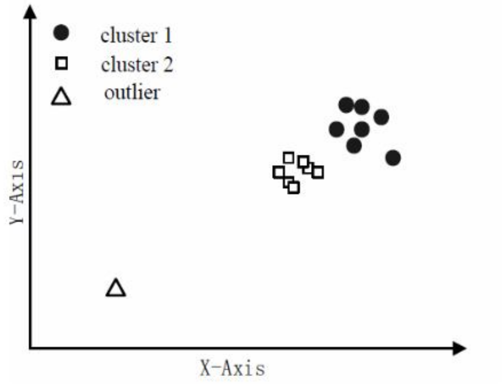


Fig.1 Samples in 2-D

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 k-means 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.

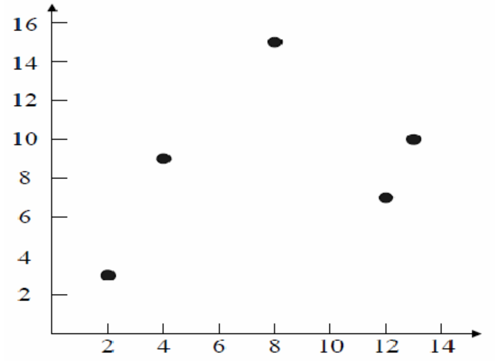


Fig.2 An outlier of samples

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. [8]. 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;

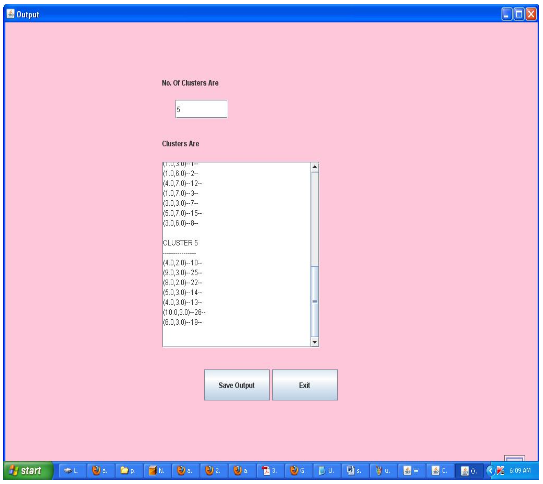
(3)In the k_{th} iteration, assuming K-1 clusters after convergence find an appropriate position for insertion of anew cluster from the set of points created in step 1 that gives minimum distortion,

(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:



No. Of Clusters Are

5

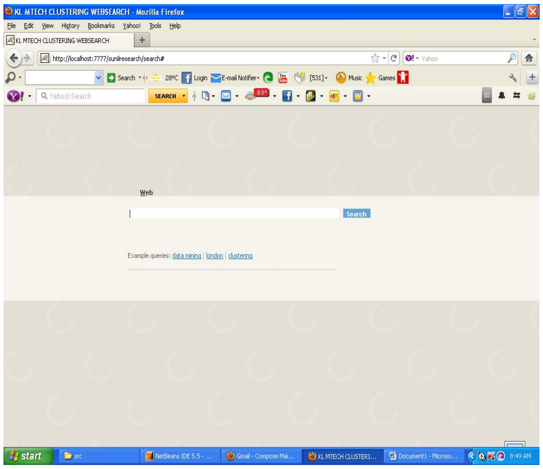
Clusters Are

```
(1.0,3.0)--1--
(1.0,6.0)--2--
(4.0,7.0)--12--
(1.0,7.0)--3--
(3.0,3.0)--7--
(5.0,7.0)--15--
(3.0,6.0)--8--

CLUSTER 5
----------------
(4.0,2.0)--10--
(9.0,3.0)--25--
(8.0,2.0)--22--
(5.0,3.0)--14--
(4.0,3.0)--13--
(10.0,3.0)--26--
(6.0,3.0)--19--
```

Fig. 3 Results showing the clustered outputs by k-means algorithm.

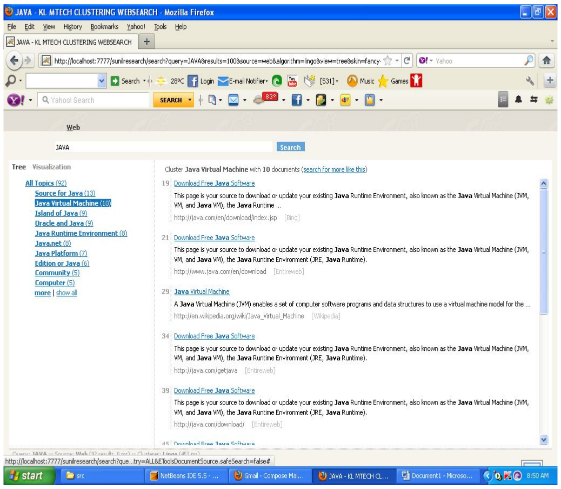
B. Online Clustering:



http://localhost:7777/sunilresearch/search#

Example queries: data mining | london | clustering

Fig. 4 Online clustering where we have to give keyword in the search text area.



http://localhost:7777/sunilresearch/search?query=JAVA&results=100&source=web&algorithm=lingo&view=tree&skin=fancy-

JAVA

All Topics (92)
Source for Java (13)
Java Virtual Machine (10)
Island of Java (9)
Oracle and Java (9)
Java Runtime Environment (8)
Java.net (8)
Java Platform (7)
Edition or Java (6)
Community (5)
Computer (5)
more | show all

Cluster **Java Virtual Machine** with **10** documents (search for more like this)

19 Download Free **Java** Software
This page is your source to download or update your existing **Java** Runtime Environment, also known as the **Java** Virtual Machine (JVM, VM, and **Java** VM), the **Java** Runtime ...
http://java.com/en/download/index.jsp [Bing]

21 Download Free **Java** Software
This page is your source to download or update your existing **Java** Runtime Environment, also known as the **Java** Virtual Machine (JVM, VM, and **Java** VM), the **Java** Runtime Environment (JRE, **Java** Runtime).
http://www.java.com/en/download [Entireweb]

29 **Java** Virtual Machine
A **Java** Virtual Machine (JVM) enables a set of computer software programs and data structures to use a virtual machine model for the ...
http://en.wikipedia.org/wiki/Java_Virtual_Machine [Wikipedia]

34 Download Free **Java** Software
This page is your source to download or update your existing **Java** Runtime Environment, also known as the **Java** Virtual Machine (JVM, VM, and **Java** VM), the **Java** Runtime Environment (JRE, **Java** Runtime).
http://java.com/getjava [Entireweb]

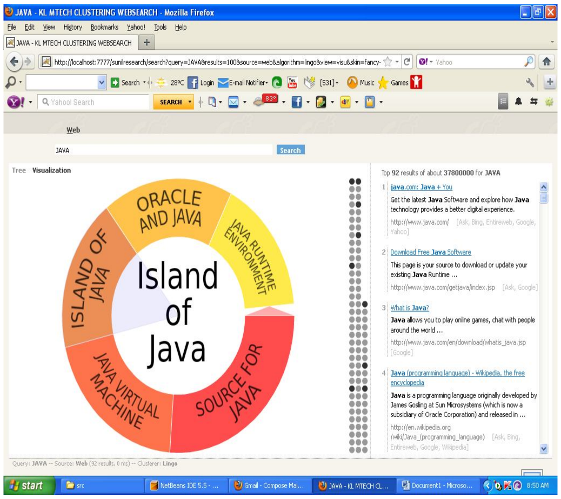
39 Download Free **Java** Software
This page is your source to download or update your existing **Java** Runtime Environment, also known as the **Java** Virtual Machine (JVM, VM, and **Java** VM), the **Java** Runtime Environment (JRE, **Java** Runtime).
http://java.com/download/ [Entireweb]

45 Download Free Java Software

Query: JAVA -- Source: Web (92 results, 0 ms) -- Clusterer: Lingo (453 ms)

http://localhost:7777/sunilresearch/search?que...try=ALL&EToolsDocumentSource.safeSearch=false#

Fig. 5 Results obtained for the given keyword ‘java’.



http://localhost:7777/sunilresearch/search?query=JAVA&results=100&source=web&algorithm=lingo&view=visu&skin=fancy-

JAVA

Top **92** results of about **37800000** for **JAVA**

1 **java**.com: **Java** + You
Get the latest **Java** Software and explore how **Java** technology provides a better digital experience.
http://www.java.com/ [Ask, Bing, Entireweb, Google, Yahoo]

2 Download Free **Java** Software
This page is your source to download or update your existing **Java** Runtime ...
http://www.java.com/getjava/index.jsp [Ask, Google]

3 What is **Java**?
Java allows you to play online games, chat with people around the world ...
http://www.java.com/en/download/whatis_java.jsp [Google]

4 **Java** (programming language) - Wikipedia, the free encyclopedia
Java is a programming language originally developed by James Gosling at Sun Microsystems (which is now a subsidiary of Oracle Corporation) and released in ...
http://en.wikipedia.org/wiki/Java_(programming_language) [Ask, Bing, Entireweb, Google, Wikipedia]

Query: JAVA -- Source: Web (92 results, 0 ms) -- Clusterer: Lingo

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

V. CONCLUSION

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. 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. If we use new clustering algorithms like DBSCAN algorithm also we can get better results because, it supports for higher data .

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 k-means 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.