International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SEARCH ENGINE INDEXING USING K-MEAN CLUSTERING TECHNIQUE

Mrs. Savita[1], Mr. Sachin Shrivastava[2]

[1]*Satya College of Engineering & Technology, Palwal, (India)*

[2]*M.tech Co-ordinator, Departement of CSE, Satya College of Eng. &Tech., Palwal, (India)*

ABSTRACT

Due to the huge growth and expansion of the World Wide Web, a large amount of information is available online. Through Search engines we can easily access this information with the help of Search engine indexing.To facilitate fast and accurate information retrieval search engine indexing collects, parses, and store data.This paper explainspartitioning clustering technique for implementing indexing phase of search engine. Clustering techniques are widely used for grouping a set of objects in such a way that objects in the same group are more to each other than to those in other groups in "Web Usage Mining". Clustering methods are largely divided into twogroups: hierarchical and partitioning methods.This paper proposesthe k-mean partitioning method of clustering and also provide a comparison of k-mean clustering and Single link HAC. Performance of these clustering techniques are compared according to the execution time based on no of clusters and no of data items being entered.

Keyword: ***Indexing, Data mining, Clustering, K-Means Clustering, Single Link HAC***

I.INTRODUCTION

Keeping in mind the end goal to encourage quick and precise data recovery, Search engine indexing gathers, parses, and stores information. As the Web continues growing, the quantity of pages filed in a web crawler increments correspondingly. With such a substantial volume of information, finding applicable data fulfilling client needs in light of basic inquiry questions turns into an inexorably troublesome errand. Internet searchers do return website pages in some rank request controlled by their match to the question, and for some situation by the significance of the pages. In any case it is regularly unthinkable for a web search tool to know which website pages might be most pertinent to the client.

The Indexing phase[1]of the web search engine is seen as a web content mining process. Beginning from the huge gathering of information the indexer first pursuit the record term which we go as a question to the web crawler likewise track the rundown of all reports that contains the given term, it additionally stores the quantity of events of every term inside each archive. This data is put away inside the list that is spoken to by utilizing Inverted document (IF) for every term t. We should store the rundown of all records that contain t and distinguish every term t by a report ID appeared as (Doc ID). We can use fixed size array for this.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

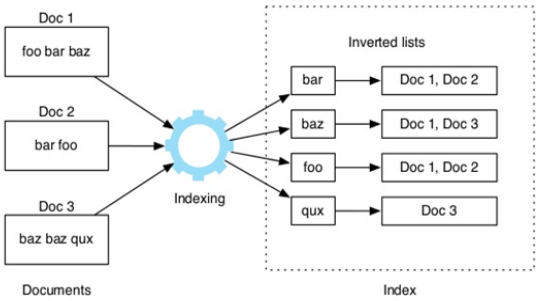


Fig.1 Indexing Process of Documents

One conceivable methodology of seeking the relevant data is to alter the positioning of returned website pages for every client so that the top positioned records all the more nearly take after the taste and enthusiasm of the client.

Document Clustering is another promising technique for sorting out list items. It is a critical information mining task. It can be depicted as the procedure of sorting out items into gatherings whose individuals are comparable somehow. Clustering can likewise be characterize as the procedure of collection the information into classes or clusters, so that the object inside a group have high similitude in contrast with each other yet are extremely not at all like items in different cluster. There are different strategies in clustering these are taken after [2]-[5].

- ❖ Partitioning Methods
- K-mean method
- K- Medoids method
- ❖ Hierarchical Methods
- Agglomerative
- Divisive
- ❖ Grid Based
- ❖ Density Based Methods
- DBSCAN
- ❖ Soft-computing Methods

Fuzzy Clustering.

This paper incorporates the K mean Partitioning [8][9] techniques for clustering. The partitioned data clustering techniques are generally classified into three main categories. They are partitioned clustering, constraint based partitioned clustering and evolutionary programming based clustering techniques.The partitioned clustering is further subdivided into two classifications which are K-Means method and other partitioned clustering algorithm like K-Mediodsmethods [10][11].

II. RELATED WORK

In the field of index organization and maintenance, many algorithms and techniques have as of now been proposed however they appear to be less productive in efficiently accessing the index. One proposed work of indexing was the threshold based clustering algorithm [8] in which the number of clusters is unknown. Be that as it may, two records are ordered to the same cluster if the similarity between them is below a specified

threshold. This threshold is characterized by the user before the algorithm begins. It is anything but difficult to see that if the threshold is small; every one of the components will get assigned to different clusters. If the threshold is large, the elements may get assigned to just one cluster. In this way the the algorithm is sensitive to specification of threshold.

Another work proposed was the Reordering algorithm In this algorithem reordering of a collection of documents D is a bijective function _: {1, . . . ,N} ! {1, . . . ,N} that maps each doc id I into a new integer _ (i). Moreover, with _ (D) = D_(1),D_(2), . . . ,D_(N). It will show the permutation of D resulting from the application of function. This algorithm is not powerful in clustering the most similar documents. The greatest document might not have likeness with any of the document but rather still it is taken as the representative of the cluster. In this technology the number of clusters is unknown. However, two documents are classified to the same cluster if the similarity between them is below a specified threshold. This threshold is characterized by the user before the algorithm begins. It is anything but difficult to see that if the threshold is small; every one of the elements will get allocated to various clusters If the threshold is large, the components may get assigned to just one cluster. In this way the algorithm is delicate to specification of threshold.

Another Work Proposed was the Agglomerative Hierarchical clustering [7] .This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the information point is recalculated yet which distance to consider when the groups has been framed? For this there are numerous accessible strategies like single-nearest distance or single linkage, Complete-farthest distance or complete linkage, average-average distance or average linkage, ward's method - sum of squared Euclidean distance is minimized. Thu-sly we continue grouping the data until one cluster is formed. Presently on the basis of dendogram graph we can compute what number of cluster ought to be really present. In any case, these methods likewise have some disadvantages as the algorithm can never fix what was done already. Time complexity of at least O(n2 log n) is required, where "n" is the number of data points. In view of the sort of distance matrix chosen for merging different algorithms can endure with one or more issue like Sensitivity to noise and outlines, Breaking vast clusters or trouble taking care of various measured clusters and convex shapes, Sometimes it is hard to recognize the right number of clusters by the dendogram.

II. PROPOSED APPROACH

The indexing Phase of the search engine gathers data from the web document assembled in the web repository. This worldwide substantial estimated record is stored in the form of inverted files. The paper has executes the k-mean algorithm to demonstrate the indexing in Search Engine and compare its execution time with Single Link HAC.

2.1 Architecture of Clustering Based Indexing in Search Engine

In this Paper architecture of indexing which has four logical phases has also been proposed. Clustering, Loading, processing and storing. Output of each phase is the input of others. Starting with clustering it take the input from repository and output of the clustering phase act as input to the loading phase. Similarly output of

loading phase, processing phase act as a input to the processing phase, storing phase respectively. Then output of the storing phase will use by the searcher to find the

Documents on the basis of user query.

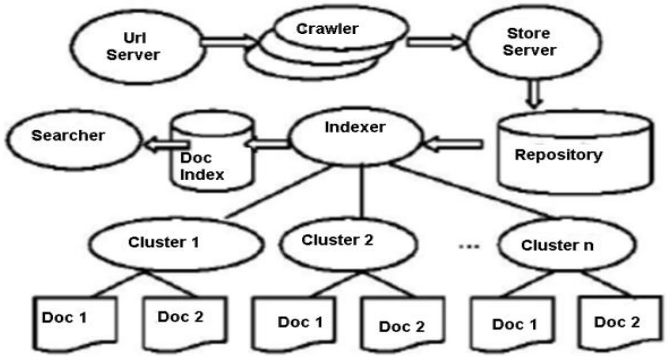


Fig 2 . Architecture of search engine Indexing

The compression of index is accomplished by applying clustering to the website pages so that the comparable pages are in the same cluster and subsequently doled out nearer identifiers. A clustering algorithm has been proposed, which converts the individual documents into k ordered clusters, and subsequently archives are reassigned new document identifiers so that the archives in the same cluster get the sequential document identifiers.

2.2 The Algorithm for Clustering

The given paper executes the K mean Clustering algorithm which is one of the most straightforward unsupervised learning algorithms that take care of the well-known clustering problem. The methodology takes after a basic and simple approach to characterize a given data set through a specific number of clusters (assume k clusters) settled earlier. The k-means clustering technique can likewise be portrayed as a centroid model as one vector representing to the mean is utilized to depict every cluster.

K-implies clustering is exceptionally helpful in exploratory data analysis and data mining in any field of research, and as the development in PC power has been trailed by a development in the event of extensive information sets. Its simplicity of usage, computational effectiveness and low memory utilization has kept the k-implies grouping exceptionally well known, even contrasted with other clustering techniques. Such other clustering techniques incorporate connectivity models like hierarchical clustering methods. These have the upside of taking into consideration an obscure number of clusters to be searched for in the data, yet are expensive computationally because of the way that they depend on the dissimilarity matrix.

A secondary objective of k-means clustering is the reduction of the complexity of the data. A decent sample would be letter grades (Faber, 1994). The numerical grades are clustered into the letters and represented by the average 16 incorporated into every class. At long last, k-means clustering can likewise be utilized as an initialization step for more computationally expensive algorithms like Learning Vector Quantization or

Gaussian Mixtures, in this way giving an estimated separation of the data as a beginning point and reducing the noise present in the dataset (Shannon, 1948)..

III. IMPLEMENTATION WORK

For indexing the documents, firstly we need to parse the documents. After that k means algorithm is applied for making the clusters. The fig. depicted the methodology of work being done in this paper.

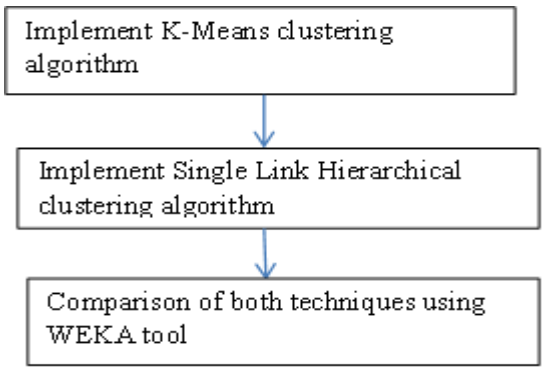


Fig.3 Methodology of Proposed Work

3.1 K-Mean Clustering Algorithm

K means clustering algorithm was developed by J. MacQueen (1967) and then by J. A. Hartigan and M. A. Wong around 1975. K-means clustering is an algorithm to arrange or to group your objects based on attributes/features into K number of group. The primary thought is to characterize k centers, one for each cluster. These centers should be set cunningly in view of various area causes diverse result. In this way, the better decision is to place them however much as could reasonably be expected far from each other. The algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c^i} (||x_i - v_i||)^2$$

Where

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j, 'c_i' is the number of data points in i^{th} cluster. 'c' is the number of cluster centers.

Algorithm: - In this we take k the number of cluster and S as data set containing items. In this output is put away as a set of k clusters. Algorithm tails some steps these are:-

Steps1:- Randomly pick k object from S as starting cluster center.

Steps2:- Calculate the distance from the data point to each cluster.

Step3: - If the data point is nearest to its own particular cluster, abandon it where it is. On the off chance that the data point is not nearest to its own cluster, move it into the nearest cluster.

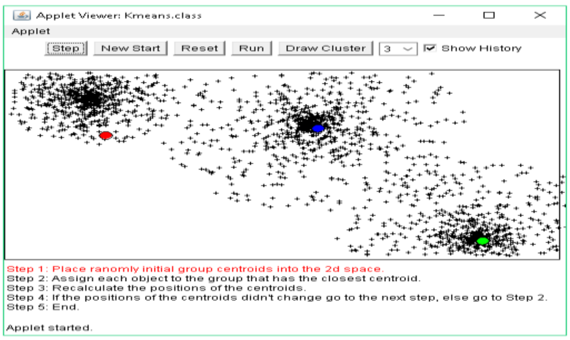
Step4: rehash step2 and 3 until best important repeat is found for every data.

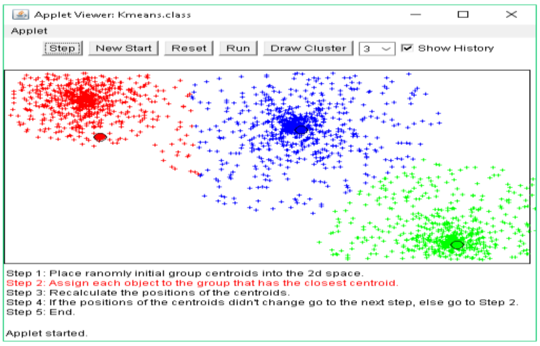
Step5: - updates the cluster means and calculate the mean estimation of the object for every object .

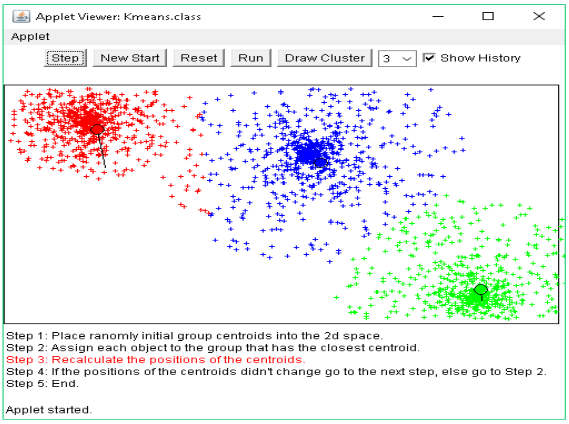
IJARSE
ISSN 2319 - 8354

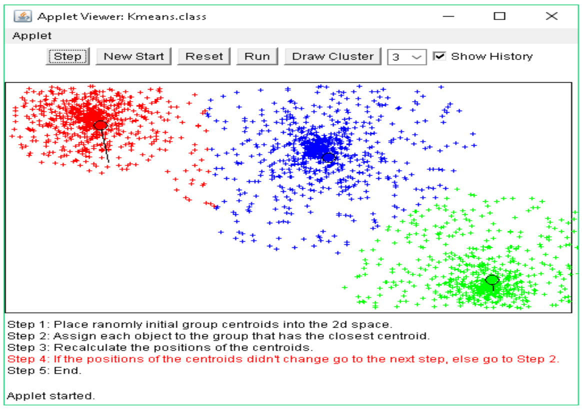
Step6: - stop (each data is situated in a proper positioned cluster).

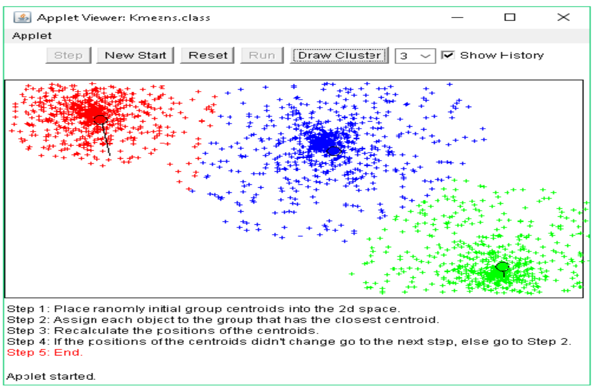
3.2 Working of K Mean algorithm using Random Data Set and Cluster Points











3.3 Single Link Hierarchal agglomerative Clustering

Single-link HAC clustering is one of the methods of Hierarchal clustering [7]. It is based on grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other.

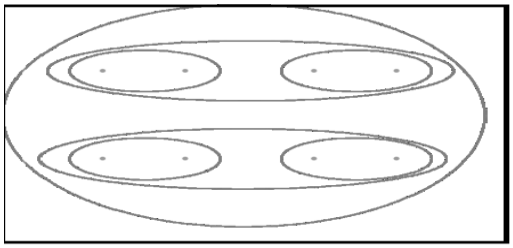


Fig 4. Single Link HAC

The single link clustering algorithm can be defined as follows :

Step: 1 Construct the distance matrix from the given pair of matrix

Step: 2 assign each pattern to a cluster

Step: 3 Use maximum similarity of pairs:

Sim (Ci, Cj) = max (Sim X, Y)

xєci, yєcj

Step: 4 Determine the smallest entry in the distance matrix D, and merge the two clusters say D (Ci , Cj)

Step: 5After merging Ci and Cj, the similarity of the resulting cluster to another cluster,Ck IsSim ((Ci U Cj), Ck) = max (Sim (CiCk),Sim (Cj,Ck))

Step:6 If only one cluster is left ,stop. Else go to step 3.

Illustration of a k-means algorithm using following data set having seven individuals:

A	B
1	1
1.5	2
3	4
5	7
3.5	5
4.5	5
3.5	4.5

This data set is to be assembled into two clusters. As an initial phase in finding a first partition, let the A & B values of the two individuals uttermost separated (utilizing the Euclidean distance measure), characterize the initial cluster means, giving:

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

The remaining individuals are now analysed in sequence and allocated to the cluster to which they are nearest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated every time another part is included. This prompts the accompanying arrangement of steps:

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Now the initial partition has changed, and the two clusters at this stage having the accompanying qualities:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

In any case, we can't yet make sure that every individual has been doled out to the right cluster. Thus, we contrast every individual's distance with its own cluster mean and to that of the opposite cluster.And we find:

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Just individual 3 is closer to the mean of the inverse cluster (Cluster 2) than its own (Cluster 1). In other words, every individual's separation to its own particular cluster mean should be smaller that the distance to the other cluster's mean (which is not the situation with individual 3). In this manner, individual 3 is moved to Cluster 2 bringing about the new partition:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

The iterative relocation would now proceed from this new partition until no more relocation happen. However, in this case every individual is currently closer its own cluster mean than that of the other cluster and the iteration stops, picking the most recent partitioning as the final cluster solution.

Likewise, it is conceivable that the k-means algorithm won't locate a final solution. For this situation it would be a smart thought to consider halting the algorithm after a pre-chosen maximum of iterations.

IV. CONCLUSION

There are different techniques accessible which can be applied for clustering; In this paper we have clarified the working of K-Mean clustering algorithm, Single Link HAC and their comparison with each other. We have evaluated algorithms on dissimilar constraints such as instances, time and number of clusters. From the values we came to realize that k-mean is better methodology when to contrast with Single Link HAC on the grounds that it requires lowly time other than Single connection HAC, distribution of cluster is also fair enough. It has greatest favorable advantage of clustering large data sets and its performance increases as number of clusters increases.

Hierarchical algorithm was adopted for categorical data, but due to its complexity a new approach for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank. As a general conclusion. The execution of K-mean calculation is superior to Hierarchical Clustering Algorithm. It is order-independent; for a given initial seed set of cluster centers, it generates the same partition of the data irrespective of the order in which the patterns are presented to the algorithm.

Comparison of both Techniques using WEKA Tool

Algorithm Implemented	No. Of Clusters	Time Taken(sec)		
		Data Set 1 Instances : 500 Attributes :6 (Avg.)	Data Set 2 Instances : 600 Attributes :12 (Avg.)	Data Set 3 Instances: 768 Attributes :9 (Avg.)
K-Mean	2	0.01	0.05	0.03
	6	0.02	0.06	0.06
	10	0.01	0.08	0.14
Single Link HAC	2	0.04	1.17	3.84
	6	0.05	1.31	3.6
	10	0.04	1.3	3.39

Table 1: Comparision of both Techniques

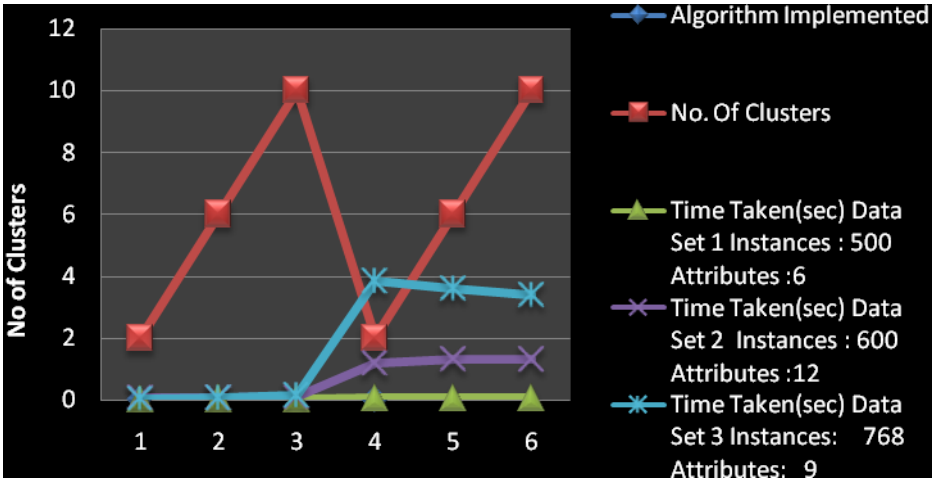


Fig 5. Execution Time

REFERENCES

[1]. FabrizioSilvestri, RaffaelePerego and Salvatore Orlando. "*Assigning Document Identifiers to Enhance Compressibility of Web Search Engines Indexes*" In the proceedings of *SAC, 2004*.

[2] M. K. Jiawei Han, Data Mining: Concepts and Techniques. Academic Press, Morgan Kaufmarm Publishers, 2001.

[3] P. Berkhin, "*Survey of clustering data mining techniques*," Springer, 2002.

[4] B. Pavel, "*A survey of clustering data mining techniques*," in Grouping Multidimensional Data. *Springer Berlin Heidelberg, 2006*, pp. 25–71.

[5] R. Xu and I. Wunsch, D., "*Survey of clustering algorithms*," Neural Networks, IEEE Transactions on, *vol. 16, no. 3, pp. 645–678,* May 2005.

[6] Ms.Aparna K1, Dr.Mydhili K Nair2 "*A Detailed Study and Analysis of different Partitional Data Clustering Techniques*", *International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization), Vol. 3, Issue 1, January 2014, ISSN: 2319-8753.*

[7] AnuradhaTyagi, Khaleel Ahmad"*Indexing in Search Engines based on Pipelining Architecture using Single Link HAC"International Journal of Computer Applications (0975 – 8887) Volume 49– No.19, July 2012*

[8]Sanjiv K. Bhatia. Adaptive K-Means Clustering. American Association for Artificial Intelligence, 2004.

[9] J. A. Hartigan and M. A. Wong, "*A k-means clustering algorithm,Applied Statistics*, 28:100--108, 1979.

[10] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data. An Introductionto Cluster Analysis, Wiley, New York, 1990.

[11] Jain, A.K., M.N. Murty and P.J. Flynn, *Data Clustering: A Review,ACM Computing Surveys, Vol. 31, No. 3, pp. 264- 323, Sep. 1999.*

[12]http://www.brickmarketing.com/define-search-engine-index.htm