Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Competent Data Set Grouping in Clustering Algorithms

Kalaivani. R*
Research scholar,
Dr. Mahalingam Centre for Research and Development
NGM College, Pollachi, India

Dr. R. Manicka Chezhian
Associate Professor
Dr. Mahalingam Centre for Research and Development
NGM College, Pollachi, India

Abstract— Clustering is a technique in which the data objects is classified into subsets or clusters, which means we can discriminate clearly whether an object belong to cluster or not. In clustering the objects of similar properties are placed in one class of objects and a single access makes the entire class available. Clustering is a division of data into groups and it corresponds to hidden patterns. It plays an outstanding role in data mining which divides the data into similar objects and models data by its cluster. Each group called cluster consist similar object between themselves and dissimilar to objects of other groups. This paper presents an overview of the methodologies and implementation in clustering either web user or web documents and presents a survey in clustering employed over web.

Keywords— Clustering Algorithm, Dempster-Shafer Theory, Artificial Ant Colony Clustering, Web Community Mining

I. INTRODUCTION

Data mining deals with large data bases that impose on clustering analysis. These led to the emergence of powerful applicable data mining clustering methods. Clustering is performed on numerical categorical attribute, subset of attribute called a segment. Elementary segment is a unit whose sub-ranges consist of a single category value or of a small numerical bin. The number of data points per every unit represents an extreme case of clustering. Segmentation is another commonly used practice in data exploration that utilizes expert knowledge regarding the importance of certain sub-domains. Clustering has been used in statistics, science, pattern recognition framework, speech and character recognition, image segmentation, computer vision data compression in image processing. Clustering in data mining was brought to life by intense development in information retrieval and text mining spatial database applications (astronomical data).sequence and heterogeneous data analysis, web applications, Deoxyribo Nucleic Acid (DNA) analysis in computational biology. Clustering resulted in a large amount of application-specific development that is beyond our scope. It plays a major role in applications such as scientific, data exploration, text mining, and information retrieval, spatial database application, marketing, medical diagnosis web analysis, Computational biology and many other.

II. RELATED WORK

Clustering Algorithm, clustering includes various methods hierarchical methods, partitioning methods, grid-based methods, [5] co-occurrence of categorical data, constraint based clustering. Clustering algorithms are used in machine Learning and high dimensional data. Hierarchical algorithm build clusters gradually, they try to discover clusters by iteratively relocating points between subsets and to identify clusters as areas highly populated with data. Probabilistic cluster include algorithm, SNOB, AUTOCLASS, MCLUST, EM framework K-Medoids methods, include algorithms PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications).

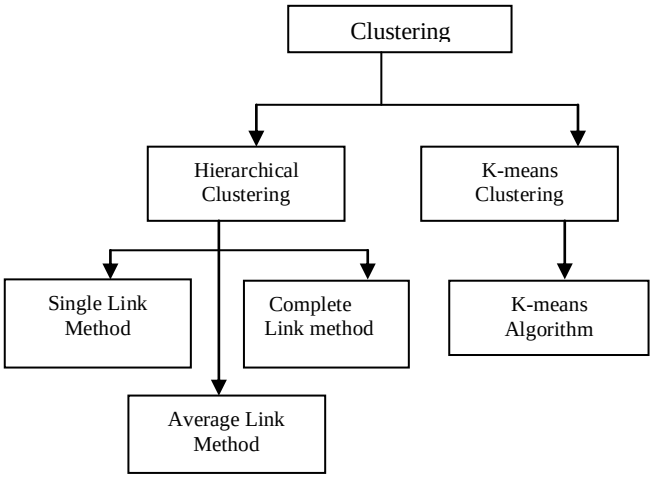

Fig 1: Clustering overview

CLARANS (Clustering Large Application based upon Randomized Search).K-means methods include initialization, optimization. Grid based methods work indirectly, they use hierarchical agglomerative as one phase of processing. Grid based method [1] include algorithms BANG, STING, Wave Clusters and this methodology also used as intermediate step in CLIQUE,MAFIA algorithms. Categorical data is connected with transaction data base. Similarity is not sufficient for clustering such data. Idea of categorical data co-occurrence is used for making cluster. ROCK (RObust hierarchical Clustering with LinKs), SNN and CACTUS (Clustering Categorical Data Using Summaries) are categorical clustering algorithms. Constraint based clustering is influence by real world data which are high dimensional data. High dimensionality reduction uses attribute transformation which is an approach of clustering attributes in groups.

III. PROPERTIES OF CLUSTERING ALGORITHMS

Clustering algorithm can handle scalability to large data set, work with high dimensional data, and find clusters of irregular shape and to handle outliers, Time Complexity, Data order dependency, Labeling or assignment, Interpretability of result. Yunjuan xie and vir.v.Phoha propose [10] dempster-shafer theory for web user clustering. Clustering and common user profile analysis using dempster-shafer theory includes extracting content pages form access log. Critical step in effective web mining is the data pre-process which include access log cleaning, session identification, low support page filtering. Access log cleaning remove the irrelevant items by checking suffix of (Uniform Resource Locator) URL request. Session identification identifies a set of user sessions by a maximal elapsed time. If the time between pages request exceeds a certain limits, we assume that the user is staring a new session use 30 minutes as a default time out. This uses greedy clustering belief function. Dempster's rule of combination to get the common user profile:

$$m_1 \oplus m_2 \ldots\ldots\ldots \oplus m_n$$

$$\text{If } 1- \sum A_i \cap B_{j=} \Phi m_1 (A_i)\ m2\ (B_j) = 0$$

m_1 and m_2 are said to be incompatible and undefined.

Abraham and Ramos proposed web usage mining [3] using artificial ant colony clustering and genetic programming. The basic mechanism underlying this type aggregation is an attraction between data items mediated by the ant workers. Small cluster of items grow by attracting workers to deposit more items. It uses Optimization technique for mining useful information. The ant clustering algorithm helps to improve performance of LGP model. Different algorithm will be also investigated to improve the trend analysis and knowledge discovery.

Yanchun Zhang and Guandong Xu investigates [9] using web clustering for web community mining and analysis. They propose web clustering algorithm for finding the linking relevant page groups through linkage structure analysis. They have combined the latent semantic analysis and web clustering to identify user session aggregation in the form of web access patterns. The propose methods which could reveal web access pattern in a latent semantic space explicitly.

Table 1: Clustering Paradigm

Clustering Methods	**Algorithms**
Hierarchical Clustering method. Agglomerative Divisive	DBSCAN(Density Based ClusteringofApplications of Noise) BRICH(Balanced Iterative Reducing and Clustering using Hierarchies) CURE (Clustering Using Representatives).
Partitioning Clustering Method.	Relocation algorithm, probabilistic clustering algorithm,k-medoid,k-means,Density based algorithm, PAM(Partition Around Medoid),CLARA (Clustering Large Applications,CLARANS(Clustering Large Applications based on Randomized Search.
Categorical Clustering method	STIRR (Sieving Trough Iterated Relational Reinforcement) ROCK(Robust hierarchical Clustering with Links) CACTUS (Clustering Categorical Data Using Summaries).

Manikandan have proposed the method for improving [4] the efficiency of textual static web content mining using clustering technique. K-Means clustering algorithm is applied to discover clusters on the document. Number of clusters depends upon the number of concepts. The clusters defines k-Centroids one for each cluster. The centroids have to be selected carefully-means algorithm improves the performance of mining process together with the generalized pattern algorithm. Generalized pattern algorithms allow users to extract small set of useful rules instead of generating a large set of trivial once. This paper proposes approach for discovering knowledge from web content. Combinations of clustering and Pattern algorithm reduced the pattern redundancy and improve the performance. Clustering is applied to group similar text documents, so that scalability is improved. Thus when a clustering input is applied efficiency of algorithm is improved.

Anjali B.Raut and G.R Bamnote propose a technique called retrieval of web documents using [2] a fuzzy hierarchical clustering to create the clusters of web documents. Two methods of clustering is used to categorize the documents, they are fuzzy c-means clustering method, and fuzzy equivalence clustering method. Fuzzy equivalence relation helps information retrieval in the terms of time and relevant information. The clustering of web document includes various steps; the first step is the downloaded documents and the keywords are stored in the database by the crawler.

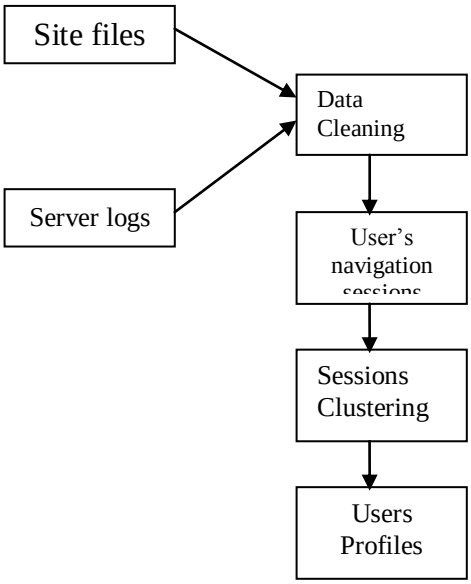

Fig: 2.Web Data Clustering

Then the indexer extracts all the words from the documents and eliminates stop words such as "a","and","the" from each document. The keywords formed fetch the related documents and stores it in the indexed database. Then the fuzzy clustering is applied on the indexed database. The fuzzy method uses fuzzy compatibility relation in terms of distance function as given below:

$$R(x_i, x_k) = 1 - \delta\left(\sum_{J=1}^{P} |x_{ij} - x_{kj}|^q\right)^{1/q} \quad \text{.....(i)}$$

Selvakumar proposed the partition clustering method for categorical [7] attribute using k-medoid algorithm. Partition clustering directly decomposes the data set into set of disjoint clusters. This offered an overview of the main characteristics of the clustering to find the better partitioning algorithm. In k-means algorithm, the number of k of clusters is fixed before the algorithm is run. The basic version begins by randomly picking k-cluster centers, assigning each point to the cluster whose mean is closest in the Euclidean distance sense and then computing the mean vectors of the points assigned to each cluster .k-means and k-medoid both attempt to minimize squared error. In contrast to k-means algorithm k-medoid chooses data points as centers. K-medoid is a classical partitioning technique of clustering. It is more robust to noise and outliers as compared to k-means. Medoid is an object of a cluster whose average is dissimilar to the entire object is minimal. The clustering algorithms k-means and k-medoid for numerical data are used to calculate accurate clustering of transactional data (real industry-commerce intelligence).k-means and k-medoid are distance based approaches that are effective for low dimensional numerical data.

IV. CONCLUSION

This work presents clustering methods and algorithm proposed by various researchers. Web mining operations include clustering, association and sequential analysis. Typical clustering operations in web mining include finding natural groupings of web resources or web users. The clusters in web mining do not necessarily have crisp boundaries.

Clustering of web documents help to discover groups of pages with related content. Web document is a collection of web pages which include HTML files, XML files, images, applets and multimedia resources. The contributions of clustering in web documents are to improve both web information retrieval and content delivery on the web.

REFERENCES

[1] Ankita Dubey, Aastha Sukrit,"Web data mining using clustering algorithm", International Journal of Engineering Technology and Management Research,vol-1,pp-42-47,2013.
[2] AnjaliB.Raut, Bamnote,"Web document clustering using fuzzy equivalence relations", Journal of emerging trends in computing and information sciences, vol-2, pp-22-27, 2010-2011.
[3] Ajith Abraham, Victorino Ramos "web usage mining using ant colony clustering and genetic programming" Oklahoma State University,Tulsa,OK 74106,USA.
[4] R.Manikandan,"Improvingefficiency textual static web content mining using clustering techniques" Journal of theoretical and applied information technology, vol-33 no.2, pp-193-198, 2011.
[5] Pavel Berkhin, "Survey of clustering data mining techniques "Accrue Software, 1045 forest Knoll Dr., San José, CA, 95129.
[6] Santhisree.k, Dr.Damodaran.A," Clustering on web usage data using approximations and set similarities", International journal of computer applications, vol-1 no-4, pp-27-31, 2010.
[7] A.Selvakumar, "An Adaptive Partitional Clustering Method for Categorical Attribute using K-medoid" International journal of computer science and mobile computing, vol.2, pp-197-204.April 2013.
[8] Vinita Shrivastava,Neetesh Gupta,"Performance Improvement of web usage mining by using learning based k-mean clustering", International Journal of computer science and its applications.
[9] Yanchun Zhang, Guandong Xu "Using web clustering for web communities mining and analysis" International Conference on Web Intelligence and Intelligent Agent Technology,2008.
[10] Yunjuan Xie, Vir.v.Phoha "Web user clustering from access log using belief function"K-cap'01, October 22-23, 2001.

