International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Review: Techniques for Clustering of Web Usage Mining

Rupinder Kaur[1], Simarjeet Kaur[2]

[1]Research Fellow, Department of CSE, SGGSWU, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Department of CSE, SGGSWU, Fatehgarh Sahib, Punjab, India

Abstract: ***In the area of software, data mining technology has been considered as important mean for discovering patterns and trends of large amount of data. So, this approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. Data mining is a computational intelligence discipline which has emerged as a effective tool for data analysis, new KDD and good decision making. The raw and unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the work of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar. The result of the clustering method and efficiency of its domain application are generally determined through various algorithms. This paper includes various clustering algorithms which can use efficiently according to the particular application, available software hardware facilities and size of dataset.***

Keywords: Web usage mining; clustering requirements; clustering techniques and algorithms.

1. Introduction

1.1 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serves the need of web-based applications. Usage data captures an origin of web users along with their browsing behavior at a web site. Web usage mining may be classified further depending on the kind of usage data considered.

(a) Web Server data
In web server data there are user logs which are collected by the web server and typically include IP address, page reference and access time.
(b) Application Server Data
Commercial application servers have significant features to enable e-commerce applications.
(c) Application Level Data

There are types of events can be defined in an application and logging can be turned off them generating histories of these events.

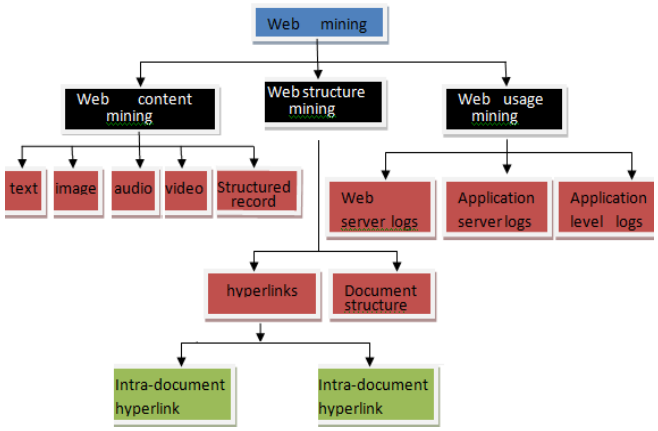

Figure 1: Web mining Taxonomy [10]

1.2Clustering or Cluster Analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering is the main task of exploratory data mining, and basically common technique for statistical data analysis used in many fields, like machine learning, pattern discovery, information retrieval and biomedical data information.

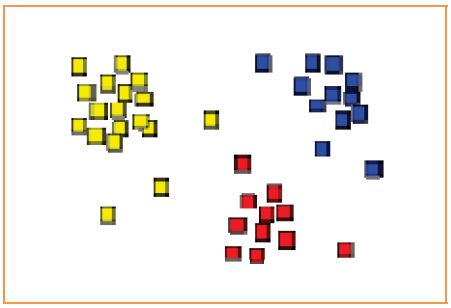

Figure 2: Clusters of Similar Objects

1.3Requirements of Clustering

The following are important requirements of cluster analysis in data mining:

1) Scalability: Some of the clustering algorithms work well on small data sets containing fewer than 200 data objects. However, a large database contains large no. of data objects.
2) Minimal requirements for domain knowledge of determine input parameters: The clustering process results can be sensitive to input parameters sometimes
3) Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based data. However, applications may also require clustering different types of data.

4) Ability to deal with noisy data: Most real-world databases contain outliners or missing, unknown, errors in data.
5) Discovery of clusters with arbitrary shape: It is needed to develop algorithms for detection clusters of arbitrary shape.
6) Insensitivity to the order of input records: Some clustering algorithms are sensitive to the order of input data; for example, may generated dramatically different clusters. Development of algorithms that are insensitive to the order of input is needed.
7) Constraint-based clustering: Real-world applications may need to perform clustering under various kinds of constraints. Suppose that you have to choose the locations for a given number of new automatic cash-dispensing machines (ATMs) in a city.
8) High dimensionality: A database or a data warehouse can contain various dimensions and attributes. Many clustering algorithms handling low-dimensional data, involving only two-three dimensions. Human eyes can judge the quality of clustering for up to three dimensions easily.
9) Interpretability and usability: Users expect clustering results should be interpretable, effective, and usable.

2. Clustering Techniques and Algorithms

2.1 Partitioning clustering technique

- Basically used to find mutually exclusive clusters of spherical.
- Based on Distance.
- Use mean or medoid to represent cluster centre.
- Usable for small and medium size data sets or objects.

2.1.1Different partioning clustering algorithms

(a) K-mean (centroid –based technique)

K-Means clustering algorithm is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points. Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects to another clusters [8].

Algorithmic steps for K-Means clustering [9]

1) *Set K* – To choose a number of desired clusters, K.
2) *Initialization* – To choose k starting points which are used as initial estimates of the cluster centroids. These cluster centriods are taken as the initial values.
3) *Classification* – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.
4) *Centroid calculation* – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.
5) *Convergence criteria* – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move.

(b) K-medoids (Object Based Technique)

In a k-medoids methods a cluster is represented by one of its points. This is an easy solution because it covers any attribute type and medoids are insensitive to outliers because peripheral cluster points do not affect them. When medoids in this algorithm are selected, clusters are known as subsets of points or values near to respective medoids, and the objective function is defined asthe averaged distance.PAM(Partitioning Around Medoids) algorithm was one of the first *k*-medoids algorithms. It is for determination *k* partitions for *n* objects. After an initial random selection of *k* representative objects, the algorithm continuously tries to make a better choice of cluster representatives. All of the possible sets or pairs of objects are analyzed, where one object in each dataset or pair is considered a representative object and the other is not. Algorithm: *k*-medoids. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Hierarchical clustering technique

- Clustering in this method is a hierarchical decomposition of dataset.
- Basically cannot correct errors of splits.
- Incorporate other techniques like micro-clustering.

2.2 Different hierarchical clustering algorithms

a) *Agglomerative clustering:* Agglomerative clustering algorithm is type of bottom-up clustering method where an agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters.
b) *Divisive clustering:* It is a top-down clustering method and is less commonly used. Divisive clustering works in a similar way to agglomerative clustering but in the opposite way. This method starts with a single cluster containing all objects, and then successively splits final clusters until only clusters of individual objects remain.
c) *Chameleon hierarchical algorithm:* Chameleon algorithm is type of hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between sets of clusters. It was derived based on the weakness of two hierarchical clustering algorithms: ROCK and CURE. Chameleon describes cluster similarity is assessed based on how well-connected objects are within a cluster *and* on the proximity of clusters. That is, two clusters are merged if connection between them is high and they are *close together*. Thus, Chameleon does not based on a static can automatically adapt to the internal characteristics of the clusters being merged. it uses a *k*-nearest-neighbor graph approach to develop a sparse graph, where vertices of the graph represents a data object, and there exists an edges between objects if one object is among the *k*-most-similar objects of the other. [3]
d) *Cure (clustering using re-presentatives):* CURE algorithm tries to handle problems of graph-based algorithm and Starts with a proximity matrix/proximity graph. It represents a cluster using multiple representative points. Main goals of algorithm are scalability, by choosing points that capture the geometry and shape of clusters and Representative points are found by selecting a constant number of points from a cluster. In this algorithm the first

representative point is chosen to be the point farthest from the center of the cluster and remaining representative points are chosen so that they are farthest from all previously chosen points.

e) *BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies:* BIRCH algorithm is designed for clustering a large amount of numerical data by integration of hierarchical clustering (at the initial *micro-clustering* stage) and other clustering methods such as iterative partitioning (at the later *macro-clustering* stage). BIRCH overcomes the two difficulties of agglomerative clustering algorithm: (1) scalability and (2) the inability to undo what was done in the previous step. BIRCH introduces two concepts, *clustering feature* and tree of feature (*CF tree*), which are used to summarize cluster representations. It helps the clustering method achieve good speed and scalability in large databases and also make it is effective for incremental and dynamic clustering of incoming objects.[3]

2.3 Density-based clustering technique

- To find clusters with arbitrary shape, density-based clustering methods have been developed.
- These regards to clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise).

2.3.1 *Several density-based clustering algorithms*

(a) *DBSCAN Algorithm*

- It stands for Density-Based Spatial Clustering of Applications with Noise.
- It is a density-based clustering technique.
- The algorithm grows regions with high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.
- DBSCAN also define cluster as a maximal set of density-connected points.

(b) *OPTICS Algorithm*

- It Stands for Ordering Points to Identify the Clustering Structure.
- OPTICS algorithm develops a set or ordering of density-based clusters.
- OPTICS constructs the different clustering simultaneously.
- The objects will be processed in a specific order.
- Due to this order selects an object that is density-reachable with respect to the lowest _value so that clusters with higher density (lower _value) will be finished first.
- Based on this idea, two values need to be stored for each object—*core-distance* and *reachability-distance*

(c) *DENCLUE Algorithm*

DENCLUE stands for Density-based Clustering .It is a clustering method based on density distribution functions. DENCLUE is built on the following key points:

- The effect of each data point can be formally modeled using a mathematical function (influence function).
- Total densities of the data space are the sum of the influence function applied to all data points.
- Clusters will be determined mathematically by identifying density attractors.

2.4Grid based clustering techniques

The grid-based clustering method uses a multi-resolution grid data structure. This algorithm quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of this method is its fast execution time, that is typically independent of the number of data objects, still dependent on only the number of cells in each dimension in the quantized space.[3]

2.4.1 Several grid based clustering algorithms

(a) STING: Statistical Information Grid algorithm

STING is a grid-based multi-resolution clustering technique in which the spatial area is divided into rectangular cells. There are several levels of such rectangular cells corresponding to different levels of resolution, and cells make hierarchical type structure i.e each cell at a high level is divided to form a no. of cells at the next lower level. Statistical information regards to the attributes in grid cells (such as the mean, max, and min values) is computed and stored. These parameters are effective for query processing.

(b) CLIQUE: A Dimension-Growth Subspace Clustering Method

CLIQUE (Clustering Inquest) was the first algorithm introduced for dimension-growth sub-space clustering. In dimension-growth subspace clustering, the clustering process will be starts at one-dimensional subspaces and grows upward to higher-dimensional ones. Because CLIQUE partitions each dimension like a grid structure and determines whether a cell is dense based on the number of points it contains, it can also be viewed as an sum of density-based and grid-based clustering methods. However, its overall approach is typical of subspace clustering for high-dimensional Space [3]

2.5 Fuzzy clustering techniques [7]

(a)Fuzzy K Means Clustering algorithm

Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees and fuzzy boundaries between clusters. Fuzzy clustering is used in fuzzy modeling, neural networks, rule-based systems.

(b)Fuzzy C Means Clustering algorithm

Bezdek [1] introduced Fuzzy C-Means clustering method in 1981, extend from Hard C-Mean clustering method. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design. FCM is mostly used in agricultural engineering, astronomy, chemistry, geology, medical diagnosis, shape analysis and recognition of target [6]. With the development of the fuzzy theory, the FCM clustering algorithm which is actually based on Ruspini Fuzzy clustering theory was proposed in 1980's. This algorithm is used for analysis based on distance between

various input data points. The clusters are produced according to the distance between data points and the cluster centers are formed for all clusters.

3. Literature Survey

Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang (2006)[11]In this paper author presented that the Fuzzy *C*-Means (FCM) is one of the algorithms for clustering based on optimizing an objective function. Due to above problem, we present the global Fuzzy *C*-Means clustering algorithm(GFCM) which is an incremental approach to clustering. It does not depend on initial conditions and the better clustering results are obtained through a deterministic global search procedure. Experiments show that the global Fuzzy *C*-Means clustering algorithm can give us more satisfactory results by escaping from the sensibility to initial value and improving the accuracy of clustering.

S.C.Punitha and M.Punithavalli (2012) [7]In this paper author presented the performance analysis of various techniques available for document clustering. The methods are grouped into three groups namely Group 1 – K-means and its variants , Group 2 - Expectation Maximization and its variants (traditional EM, SG EM algorithm and (LPR) using EM algorithms), Group 3 - Semantic-based techniques (Hybrid method and Feature-based algorithms). Several experiments were conducted to analyze the performance of the algorithm and to select the winner in terms of cluster effectiveness, clustering accuracy and speed of clustering.

K. Suresh R. Madana Mohana and A. Rama Mohan Reddy (2011) [4] In this paper author presented that clustering method is very sensitive to the initial center values ,requirements on the data set too high, and cannot handle noisy type data the this proposed method is using information entropy to initialize the cluster centers and introduce weighting parameters to adjust the location of cluster centers and noise problems. The navigation data sets are sequential, Clustering web data is finding the some groups which share common interests and behavior by analyzing the data collected in the web servers, this improves clustering on web datasets effectively using improved FCM clustering. Web usage mining is technique used to web log data repositories. It is used in finding the access patterns of user from web access log.

Soumi Ghosh and Sanjay Kumar Dubey (2013)[8]In this paper author presented that the outcome of the clustering process and efficiency of its domain application are generally determined through algorithms. There are different algorithms which are used to solve this problem. In this work two main clustering algorithms namely centroid based K-Means and representative object based FCM clustering algorithms are compared. These two algorithms are applied and throughput is evaluated on the basis of the efficiency of clustering output. The no. of dataset points as well as the no. of clusters is the factors upon which the behaviour patterns of both the algorithms are analyzed. FCM gives results like K-Means clustering but it still requires more computation time than K-Means clustering.

Mrutyunjaya Panda and Manas Ranjan Patra (2008)[5]In this paper author presented that besides the limited memory and Constraints of one-pass, the behavior of evolving data streams implies the following requirements for stream clustering: no assumption on the number of clusters, discovery of arbitrary shaped clusters and ability to handle outliers. Type of traditional instance-based learning techniques can only be used to detect known intrusions, since these methods divides instances based on learning they have. They detect some of new intrusions since these intrusion classes has not been able to detect new intrusions as well as known intrusions. In this paper, there is proposal of some clustering algorithms such as K-Means and Fuzzy c-Means for network intrusion detection.

Helo´ına Alves Arnaldo and Benjam´ın R. C. Bedregal (2013)[2] In this paper author presented that data clustering is main work in data mining, image processing and other pattern accessing problems. Most popular clustering algorithm is the Fuzzy C-Means. The performance of the FCM is strongly affected by the selection of the initial centroid clusters. Therefore, choose good set of initial centroid clusters is main thing for the algorithm. However, it is very hard to select a good set of initial centroid clusters randomly. In this paper, there is proposal of a method to obtain the initial centroid clusters in the FCM to accelerate the process of clustering and improve the quality of the clustering.

4. Conclusion

The cluster analysis examines unlabeled data, by either constructing a hierarchical structure, or forming a set of groups, according to a pre-specified number. In this paper, there is an description of basic concept of clustering, by first providing the definition of different clustering algorithms. The main focus was on these clustering algorithms, and a review of a wide variety of approaches that are mentioned in the introduction. These clustering algorithms evolve from different research communities, and these methods reveal that each of them has advantages and disadvantages. So, there are various clustering algorithms which can used efficiently according to the particular application, available software hardware facilities and size of dataset.

References

[1] Jain, R. Duin, and J. Mao, “Statistical pattern recognition: A review,” IEEE Trans. Pattern Anal. Mach.Intell., vol. 22 , 2000.
[2] Helo´ına Alves Arnaldo and Benjam´ın R. C. Bedregal“A new way to obtain the initial centroid clusters in Fuzzy C-Means algorithm” , 2013.
[3] Jiwei Han and Micheline Kamber,”Data Mining Concepts and Techniques” ,2006.
[4] K.Suresh R.Madana Mohana and A.RamaMohanReddy “ Improved FCM algorithm for Clustering on Web Usage Mining”,vol. 8,2011.

[5] Mrutyunjaya Panda and Manas Ranjan Patra"Some clustering algorithms to enhance the performance of the network intrusion detection system" ,2008.
[6] R. Davé and R. Krishnapuram, "Robust clustering methods: A unified view," IEEE Trans. Fuzzy Syst., vol. 5, May 1997.
[7] S.C.Punitha and M.Punithavalli "A Comparative Study to Find a Suitable Method for Text Doc ument clustering volno. 10,2012.
[8] Soumi Ghosh , Sanjay Kumar Dubey (Department of Computer Science and Engineering, Amity University, Uttar Pradesh, Noida, India)" Comparative Analysis of K-Means and Fuzzy C-Means Algorithms",Vol. 4, 2013.
[9] T. Kanungo and D. M. Mount, "An Efficient K-means Clustering Algorithm: Analysis and Implementation", Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 24, no. 7, 2002
[10] Wang, Yan. "Web Mining and Knowledge Discovery of Usage Patterns", 2012.
[11] Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang "The Global Fuzzy C-Means Clustering Algorithm", 2006.