A Survey on Different Clustering Techniques for Web Usage Mining

Mohammed Asad
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India.

Girish P. Potdar
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India.

***Abstract*—Web usage mining refers to the application of data mining techniques to discover useful patterns of different users from web log data. Basic aim of web usage mining is 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 of web users along with their browsing behavior at a web site. Clustering is the process of grouping data objects together in such a way that the objects within the same group are similar to each other whereas objects belonging to different groups are dissimilar. Clustering techniques are widely used in Web Usage Mining to extract similar patterns among users accessing a Web site. This paper reviews some of the popularly used clustering techniques: *k*-Means, *k*-Medoids, Leader and DBSCAN. These techniques are overviewed against the Web user navigational data.**

Keywords-web usage mining, k-means, k-medoids, DBSCAN, Leader.

I. Introduction

Data mining is nothing but the extraction of hidden predictive information from large databases. It is a technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools are used to predict future trends and behaviours from raw data which helps to make useful decisions in business. Data mining technology can answer business questions that traditionally were too time consuming to resolve. They process on databases for hidden patterns, finding predictive information that is useful to experts for decision making. Fig.1 shows a typical data mining process.

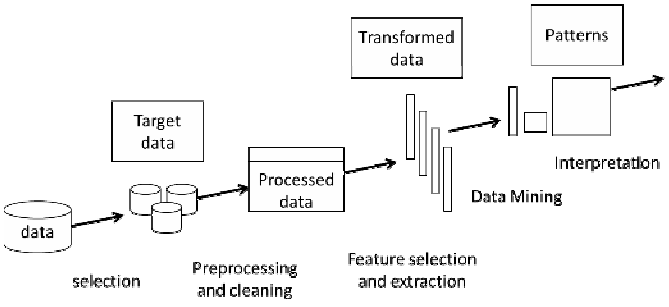

Fig.1 Data Mining Process

Web mining is one of the application of data mining techniques to extract useful information from data taken from web. Different sources of data collection in web mining are server, client, proxy servers, or obtained from an organization’s database [1]. Nature of web data is generally distributed, semi-structured, unlabelled, heterogeneous and time varying. Hence in order to handle such kind of data to extract knowledge web mining technique is used [2]. Fig.2 shows the taxonomy of web mining.

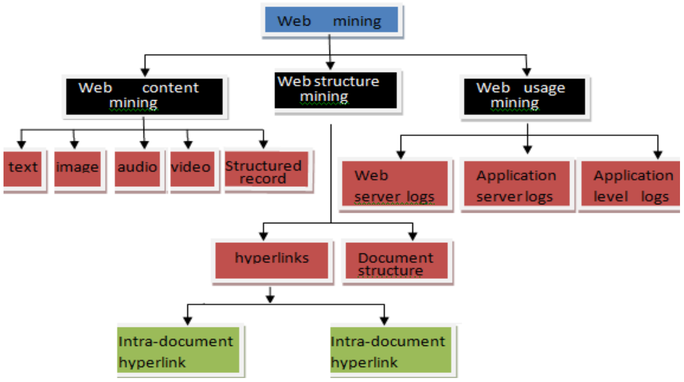

Fig. 2 Web Mining Taxonomy

Web mining can be categorized as:

- Web Content Mining deals with the extraction of knowledge from multimedia documents, consisting text, images, audio and video information. Due to this mining, the extraction of concept relations from the Web, and their automatic categorization is done.

- Web Structure Mining is done on various inter-document links, given as a graph of links in a site or between sites. For example, in Google, if several important pages point to a particular page then that page is also important.

- Web Usage Mining of the data generated by the users' access patterns of the Web, typically represented as Web server access logs, user profiles, user queries and mouse-clicks. This includes trend analysis, and Web access association/sequential pattern analysis.

A. *Web Usage Mining*

Web Usage Mining is described as the automatic discovery and analysis of patterns in web logs [3]. This web log file is generated as a result of user interactions with Web resources on one or more Web sites. The main aim of Web usage mining is to collect, analyze and extract the behavioural patterns and profiles of users interacting with a Web site. The extracted patterns are generally represented as sets of URLs that are frequently accessed by groups of users with common interests. Web usage mining has been used in a variety of applications such as i) Web Personalization systems, ii) Adaptive Web Sites, iii) Business Intelligence, iv) System Improvement to understand the web traffic behaviour which can be utilized to decide strategies for web caching, load balancing and data distribution [4] [5]. Fig.3 shows web usage mining process.

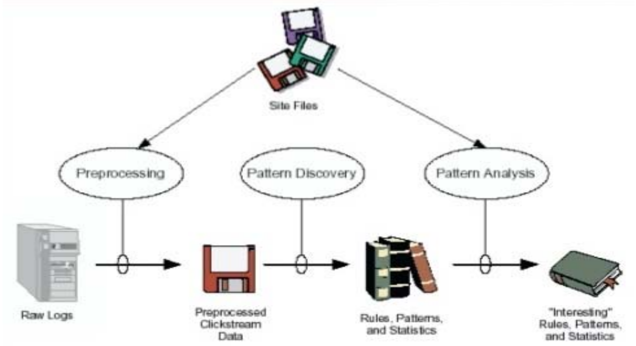
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig.3 Web usage mining process

B. *Clustering*

Clustering techniques are used in Web Usage Mining to extract similar patterns among users accessing a Web site. Clustering aims to create groups or clusters of a data set in which inter-cluster similarities are minimum and the intra cluster similarities are maximum [6]. The purpose of clustering is to allocate data points to a system of k clusters. Clustering groups the data points based on the information found in the data which describes the data objects and the relationships between them. Fig.4 shows clustering scenario.

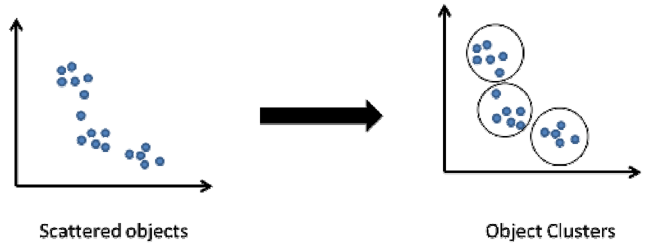

Fig.4 Clustering Scenario

Some of the major categories of the clustering techniques are: i) partitioning methods: In these methods creation of k partitions of a given data set, each representing a cluster is done [7]. Widely used partitioning methods include k-means and k-medoids. In k-means algorithm each cluster is represented by the mean value of the data points in the cluster called centroid of the cluster. On the other hand and in k-medoids algorithm, each cluster is represented by one of the data point located near the centre of the cluster called medoids of the cluster. Leader clustering is also a partitioning based clustering techniques which generates the clusters based on an initially calculated dissimilarity measure, ii) Hierarchical methods: In these methods creation of hierarchical decomposition of the set of data objects is done [7]. A hierarchical method can be categorized as agglomerative or divisive, this classification depends upon the formation of hierarchical decomposition. iii) Density- based methods: Formation of clusters based on the concept of density. Clusters of arbitrary shapes can be discovered by them. These methods continue growing clusters unless the number of objects or data points in the "neighbourhood" exceeds some threshold. DBSCAN is nothing but a density-based method in which cluster sizes grow with respect to a density-based connectivity analysis. iv) Grid-based methods: These methods quantize the data points into a finite number of cells that form a grid structure. On the grid structure, all the clustering operations are performed.

II. WEB USAGE MINING USING CLUSTERING

A number of clustering algorithms have been used in Web usage mining where the data items are user sessions consisting of sequence of page URLs accessed and interest scores on each URL page based on the characteristics of user behaviour such as time elapsed on a page or the bytes downloaded [8]. In the domain of web usage mining, clustering can be applied in two different ways, either to form clusters of users or to form clusters of data items. In user oriented clustering, different users are grouped together on the basis of similarity of their web page access patterns. In data item oriented clustering, data items are grouped together on the basis of similarity of the interest scores for these items across all users [9].

III. Clustering Techniques

A. K-means Clustering Algorithm

The k-Means clustering algorithm is a widely used clustering techniques for partitioning the data. Given a set of *m* data points , where each data point is an *n*-dimensional vector, *k*-means clustering algorithm aims to partition the *m* data points into *k* clusters ($k \leq m$) $\boldsymbol{C} = \{c_1, c_2, \ldots, c_k\}$ so as to minimize an objective function $J(V, X)$ of dissimilarity, which is the within-cluster sum of squares [10]. In most of the times measurement of dissimilarity is done based on the Euclidean distance. The objective function is nothing but to indicate the distance of the *n* data points from their respective cluster centres. The objective function *J* between a data point x_i in cluster *j* and the corresponding cluster centre v_j, is defined as.

$$J\,(X,V) = \sum_{j=1}^{k} \sum_{i=1}^{m} \mu_{ij}\, d^2\,(x_i, v_j) \qquad (1)$$

Where,

$d^2\,(x_i, v_j)$ is the Euclidean distance between x_i and v_j

μ_{ij}=1 if x_i Є c_j and 0 otherwise

At the very beginning in the k-means clustering, initialization of the cluster centers is done. This process is done randomly. After that every data object x_i is connected to one of the clusters v_j which has the minimum distance with this data point. Once all the data points have been assigned to clusters, cluster centers are updated by taking the weighted average of all data points in that cluster. The process is continued until there is no change in cluster centers. The partitioned clusters are defined by an $m{\times}k$ binary membership matrix $\boldsymbol{U}$, where the element u_{ij} is 1, if the i^{th} data point x_i belongs to the cluster *j*, and 0 otherwise [11].

The k-means algorithm gives optimum outputs in context of the sum of squared errors represented by the error objective function. As this algorithm is fast and iterative in nature, it has been applied to a variety of areas.

K-means algorithm attracts various experts due to its simplicity and flexibility. On the other hand, it has some major disadvantages due to which it is not being implemented on large datasets [12]. The most important among these are i) k-means scales poorly with respect to the time it takes for large number of points; ii) The algorithm might converge to a solution that is a local minimum of the objective function. The major drawback of k-means algorithm is its sensitivity to select the initial cluster centres.

B. K-medoids Clustering Algorithm

K-medoid is another partitioning based clustering in which *m* data points are grouped together into *k* clusters. This technique tries to minimize the squared error, which is defined as the distance between data objects within a cluster and a point designated as the center of that cluster. K-Medoids algorithm differs from k-means algorithm in a way that it selects data objects as cluster centers (or medoids). A medoid can be defined as the data point in a cluster, whose average dissimilarity to all the other data points in the cluster is minimal i.e. a medoid is a most centrally located data object in the cluster.

Given a set of *m* data points $X = \{x_i \mid i=1 \ldots m\}$, where each data point is a *n*-dimensional vector, k-medoids clustering algorithm aims to partition the *m* data points into *k* clusters ($k \leq m$) $\boldsymbol{C} = \{c_1, c_2, \ldots, c_k\}$ for minimizing the objective function which represents the sum of the dissimilarities between each of the data points and its corresponding cluster medoid [13]. Let $\boldsymbol{M} = \{m_1, m_2, \ldots, m_k\}$ be the set of medoids corresponding to *C*. The objective function $J(X, M)$ is defined as

$$J\,(X\,,M) = \sum_{j=1}^{k} \sum_{i=1}^{m} \mu_{ij}\, d^2\,(x_i, m_j) \qquad (2)$$

Where,
x_i is the i^{th} data point
m_j is the medoids of cluster c_j
μ_{ij}=1 if x_i Є c_j and 0 otherwise
$d^2\,(x_i, m_j)$ is the Euclidean distance between x_i and m_j

The basic idea behind using k-Medoids clustering algorithms is to discover *k* clusters in *m* objects. First task is to select a data point randomly i.e. medoid of that cluster. Now every other data object is assigned to a particular cluster with the medoid to which it is the most similar. *K* is the input given to algorithm, which is the number of clusters to be partitioned among a set of *m* objects.

K-medoid is more robust as compared to k-means for noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. In it minimization of sum of pair-wise dissimilarities is done instead of a sum of squared Euclidean distances as in case of k-means.

C. DBSCAN Clustering Algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based data clustering algorithm because it generates the number of clusters from the calculated density distribution of related nodes [15].

In this algorithm “ε-neighbourhood” and “density reachability” notions are used for creating clusters. Let the ε-neighbourhood of a data point x_p, denoted as is defined as below

$$N \square\ (x_p)\ \{x_q \in X \mid d^2 (x_p,x_q) \leq \square\ \} \qquad (3)$$

Where,
□ is the neighbour distance

x_q is called density-reachable from x_p if there is a sequence x_1 , … , x_n of points with $x_1 = x_p$ and $x_n = x_q$ where each $xi + 1$ is directly density-reachable from x_i. Two data points x_p and x_q are treated as density-connected if there is a point x_o such that x_o and x_p as well as x_o and x_q are density-reachable.

A cluster of data points satisfies two properties: i) All the data points in a cluster are mutually density-connected. ii) If a data point is density-connected to any data point of a cluster, it is part of that cluster as well.

Input to DBSCAN algorithm are i) ε (epsilon) and ii) η, the minimum number of points required to form a cluster. The algorithm is started by random selection of an initial data point that has not been visited. If the ε-neighborhood of this data point contains sufficiently many points, a cluster is started. Else, the data point is treated as noise. Later this point might be found in a sufficiently sized ε-neighborhood of a different data point and hence could become part of a cluster. If a data point is found to be part of a cluster, all the data points in its ε-neighborhood are also part of that cluster and hence added to the cluster. This process grows continuously until the cluster is completely created. After that, a new unvisited point is taken and previous operations are performed, leading to the discovery of a next cluster or noise.

D. Leader Clustering Algorithm

The leader clustering algorithm is based on a predefined dissimilarity threshold [14]. Initially, from the input data set, a data point is selected randomly and treated as leader. Then, distance of every other data point with the selected leader is calculated. If the distance of a data point from the leader is less than the specified dissimilarity threshold that data point falls in the cluster with the initial leader. Else, that data point is recognized as a new leader. The computation of leaders is continued until all the data points are covered. It should be considered that the quality of clusters depends on the specified distance threshold. The selected threshold is inversely proportional to the number of leaders.

Given a set of m data points $X = \{x_i \mid i=1 \ldots m\}$, where each data point is an n-dimensional vector. The Euclidean distance between the i^{th} data point $x_i \in X$ and j^{th} leader $l_j \in L$ (where L is a set of leaders) is given by

$$d^2 (x_i, l_j) = \left| \sum_{k=1}^{n} x^i{}_k - l^j{}_k \right|^2 \qquad (4)$$

Where,
n is the number of dimensions of each data point
$X^i{}_k$ is the value of k^{th} dimensions of x_i
$l^j{}_k$ is the value of k^{th} dimension of x_j

Although Leader clustering algorithm does not require estimating the value of k at the beginning, it does require estimating the dissimilarity threshold ε. Number of clusters formed in Leader clustering is inversely proportional to the value of dissimilarity threshold ε.

CONCLUSION

The focus of this paper is describing different clustering techniques of web usage mining processes. After describing the task of clustering, the most common clustering methods such as k-Means, k-Medoids, Leader and DBSCAN clustering algorithms were enumerated based on their fundamental approach. These algorithms serve as basis for the web usage clustering.

REFERENCES

[1] Kosala and Blockeel, "Web mining research: A survey," SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, 2000.

[2] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining," in Data Warehousing and Knowledge Discovery, pp. 303– 312, 1999.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan. "Web usage mining: Discovery and applications of usage patterns from web data.", SIGKDD explorations, pp. 12–23, 2000.

[4] B. Mobasher, "Data mining for web personalization." Lecture Notes in Computer Science, 2007.

[5] Etzioni O. Perkowitz, M., "Adaptive web sites: Automatically synthesizing web pages.", In Proceedings of the 15th National Conference on Artificial Intelligence, Madison, pp. 727-732, 1998.

[6] Ajith Abraham., "Business intelligence from web usage mining.", Journal of Information & Knowledge Management, pp. 375–390, 2003.

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

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

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

[10] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa., "Discovery and evaluation of aggregate usage profiles for web personalization.", Data Mining and Knowledge Discovery, pp. 61–82, 2002.

[11] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl., "An algorithmic framework for performing collaborative filtering.", In Proceedings of the 22nd annual international ACM SIGIR conference on

Research and development in information retrieval, SIGIR '99, pp. 230–237, 1999.

[12] L. Kaufman, P.J. Rousseeuw, "Finding Groups in Data. An Introduction to Cluster Analysis", Wiley, New York, 1990.

[13] Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos., "Exploiting web log mining for web cache enhancement.", In WEBKDD 2001 Mining Web Log Data Aacross All Customers Touch Points, volume 2356 of Lecture Notes in Computer Science, pp. 235–241, 2002.

[14] T. R. Babu, M.N. Murty, "Comparison of Genetic algorithm based prototype selection scheme", Pattern Recognition pp. 523–525, 2001.

[15] Zahid Ansari, Mohammed Fazle Azeem, A. Vinaya Babu, and Waseem Ahmed., "Preprocessing users web page navigational data to discover usage patterns.", In The Seventh International Conference on Computing and Information Technology, Bangkok, Thailand, May 2011.