ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
*Vol. 1, Issue 4, June 2013*

# A Study of Web Log Analysis Using Clustering Techniques

Hemanshu Rana[1], Mayank Patel[2]

Assistant Professor, Dept of CSE, M.G Institute of Technical Education, Gujarat India[1]

Assistant Professor, Dept of CSE, M.G Institute of Technical Education, Gujarat India[2]

**Abstract:** Web usage mining is the area of web mining which deals with the extraction of interesting knowledge from web log information produced by web servers. Web usage mining techniques can be applied for web log analysis. Web access data, traditionally, are stored in the server log files. Several web usage mining approaches have been presented for exposing usage patterns with the most prominent ones being clustering, association rule, and sequential pattern mining. In this paper, three different algorithms are reviewed for generating clusters. The first one is simple K-means, second K-means using Neural Network concept and Self Organization Map (SOM). , This paper deals with study of a two-stage method that integrates algorithms, first of which uses Self-Organizing Feature Maps neural network to determine the number of clusters and cluster centroids, then the second one is a K-means algorithm to find the final solution.

**Key words*:*** Web-log analysis, Clustering, K-means, SOM, Neural Network.

## I. INTRODUCTION

Web logs, generally referred to as *blogs*, are considered as online diaries published and maintained by individual users (bloggers), bloggers' daily activities reports. Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. This paper provides survey of web usage mining including web log analysis using clustering. Section 2 describes web data that can be useful for web usage mining. Section 3 describes review of existing efforts in web usage mining, web log data preparation & related work on mining of web access logs. Sections 4 describe various clustering techniques. Section 5 presents experiment results, and finally section 6 draws the conclusion.

## II. WEB DATA

Knowledge discovery in databases is to create a suitable target data set for the data mining tasks. In web mining, data can be collected at the server-side, client-side, and proxy-servers or obtained from an organization's database. There are many kinds of data that can be used in web mining, such as content, structure, usage, and user profile.

- Content: Actual data in web pages.
- Structure: This type of data describes the organization of the content. Intra page structure information includes the arrangement of various HTML or XML tags within a given page, which is generally represented as a tree structure. Main principle of intra page structure is hyper links connecting one page to another.
- Usage: Data that describes the pattern of usage of web pages, like IP address, page reference, and date and time of access.
- User Profile: Data that provide demographics information about user of the web site. This includes registration data and customer profile information.

### *A. Data Sources and Abstractions*

Web data are collected at different levels such as server level collection, client level collection, and proxy level collection. The information provided by these data sources can be used to construct/identify several data abstractions, notably users, server sessions, episodes, click-streams, and page views.

## III. WEB USAGE MINING

There are three main tasks for performing web usage mining or web usage analysis, which are shown in Fig. 1. This section presents an overview of the tasks for each step and discusses the challenges involved.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
*Vol. 1, Issue 4, June 2013*

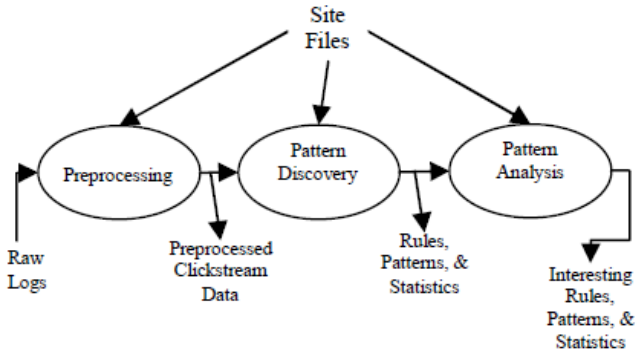


Fig:-1 High Level Web Usage Mining Process [11]

The *preprocessing stage* includes processing on all three types of data – usage preprocessing, content preprocessing and structure preprocessing. This step also includes steps such as data cleaning, efficient user identification, session identification and path completion, and transaction identification.

The *pattern discovery* can be obtained by different techniques some of which are statistical analysis, association rules, clustering, classification, sequential patterns, and dependency modeling.

The *pattern analysis* is the last step in the overall web usage mining process as described in Figure 1. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations.

*A. Related Work on Mining of Web Access Logs*

Many data preparation techniques are presented by [3, 4] in order to perform web usage mining on web access logs. However, there is not much literature available that describes preprocessing in detail for web usage mining [1]. The problems involved in web usage mining have been discussed in [3, 7]. One of the problems in getting an accurate picture of the website access is caused by web browsers and proxy servers. Web browsers store pages that have been visited and if the same page is requested, the web browser displays the page rather than sending another request to the web server. Proxy server's cache frequently visited pages locally to reduce network traffic and improve server performance.

## IV. CLUSTERING TECHNIQUES

Clustering is an unsupervised classification technique widely used for web usage mining with main objective to group a given collection of unlabeled objects into meaningful clusters [2]. For the web domain the objects are either web documents, or references of web documents, or user visits. Different techniques (algorithms) are available in order to perform clustering. These algorithms can be categorized into following major and minor subgroups:

- *Partitioning Algorithm*
  - ✓ *K-means*
  - ✓ *Neural Network*
  - ✓ *Genetic Algorithm*
- *Hierarchical Algorithms*
- *Density base Algorithms*
- *Model base Algorithms*

Traditional clustering uses the following characterizations for the derived clusters [14]:

- *Exclusive* or *hard* clusters, when an object belongs to one and only one cluster;
- *Overlapping* clusters, when an object may belong to several clusters, while no additional information is provided about the membership of the objects into the appropriate clusters;
- *Probabilistic* clusters, where an object belongs to any cluster with a certain probability; and
- *Fuzzy* clusters, where an object belongs to each cluster with a degree of membership.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
*Vol. 1, Issue 4, June 2013*

*A. Clustering K-means*

Data clustering is one of the fundamental operations in data mining, machine learning and pattern classification applications. Clustering in N-dimensional Euclidean space is the process of partitioning a given set of *n* points into a group of *k* clusters) based on some similarity (distance) metric. Generally, the Euclidean distance (Eq. 1), which derived from the Minkowski metric, is considered for similarity measure.

$$d(x,y) = \sqrt{\sum_{i=i}^{m}(xi - yj)^2} \ldots\ldots(1)$$

The existing clustering algorithms can be simply classified into the following two categories: hierarchical clustering and partition clustering. The most class of popular class of partition clustering methods is the center based clustering algorithms. The k-means algorithms, is one of the most widely used center based clustering algorithms. To find *K* centers, the problem is defined as an optimization (minimization) of a performance function, Perf(X, C), defined on both the data items and the center locations. A popular performance function for measuring goodness of the k clustering is the total within-cluster variance or the total mean-square quantization error (MSE), as shown in Eq. 2 as,

$$Perf(X,C) = \sum_{i=1}^{N} \min\{\|xi - cl\|, l = 1,..,k\} \ldots\ldots(2)$$

The k-means algorithm is by far the most popular clustering tool used in web log analysis. The name comes from representing each of *k* clusters $C_j$ by the mean (or weighted average) $c_j$ of its points, the so-called centroid. While this obviously does not work well with categorical attributes, it has the good geometric and statistical sense for numerical attributes. The sum of discrepancies between a point and its centroid expressed through appropriate distance is used as the objective function.

The steps of the k-means algorithm are as follow [8]:
*Step 1:* Choose *K* cluster centers randomly from *n* points
*Step 2:* Assign each point to clusters
*Step 3:* Compute new cluster centers
*Step4:* If termination criteria satisfied, stop otherwise continues from step 2
Note that in case the process close not terminates at step 4 normally, then it executed for a mutation fixed number of iterations.

*B. Self Organization Map*

The Self-Organizing Feature Map (SOM) [8, 12], which is typical of the unsupervised learning neural networks, can project a high-dimensional input space on a low-dimensional topology so as to clusters directly. Teuvo Kohonen [12] introduced the SOM network that reduced the dimensions of data through the use of self organizing neural networks. The SOM network produces a map of usually one or two dimensions which plot the similarities of the data by grouping similar data items together. This mapping process reduces the problem dimensions. The SOM network integrates dimensions reducing and clustering in one network. Figure 2 shows the mapping from a one-dimensional input to a two-dimensional array. The SOM network organizes itself by competing representation of the samples. Neurons are also allowed to change themselves in hoping to win the next competition. This selection and learning process makes the weights to organize themselves into a map representing similarities.

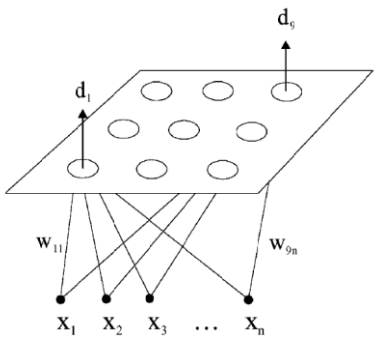


Fig:2 The Mapping from a one-dimensional input to a two-dimensional array [11].

The most widely used unsupervised learning scheme is the self-organizing feature maps. Weight calculation using

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
*Vol. 1, Issue 4, June 2013*

following equation:

$$Wij=Wij+\eta\,(Xij-Wij) \text{ ..... } (1)$$

where η is learning rate, Xij is data, Wij is weight.

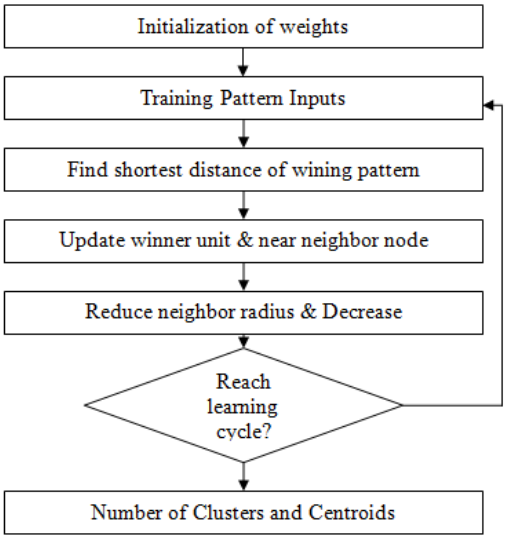


Fig:3 SOM Algorithm [8]

*C. K-means using Neural Network*

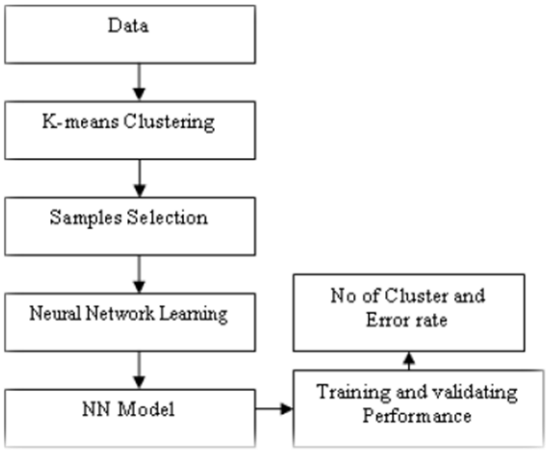


Fig:-4 K-means using NN Algorithm

In this section, a method combining neural network with K-means algorithm as proposed in [13] is reviewed to perform web log analysis. Here, the role of the k-means algorithm is to reduce the computation intensity of the neural network, by reducing the input set of samples to be learned. This can be achieved by clustering the input dataset using the k-means algorithm, and then take only discriminate samples from the resulting clustering schema to perform the learning process. By doing so, we are trying to select a set of samples that cover at maximum the region of each class in the N-dimensional space (N is the size of the training vectors). The input classes are clustered separately in such a way to produce a new dataset composed with the centroid of each cluster, and a set of boundary samples selected according to their distance from the centroid. Reducing the number of used samples will enhance significantly the learning performances, and reduce the training time and space requirement, without great loss of the information handled by the resulting set, due to its specific distribution. The number of fixed clusters (the k parameter) can be varied to specify the

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
*Vol. 1, Issue 4, June 2013*

coverage repartition of the samples. The number of selected samples for each class is also a parameter of the selection algorithm. Then, for each class, we specify the number of samples to be selected according to the class size. When the clustering is achieved, samples are taken from the different obtained clusters according to their relative intra-class variance and their density (the percentage of samples belonging to the cluster). The two measurements are combined to compute a coverage factor for each cluster. The number of samples taken from a given cluster is proportional to the computed coverage factor.

## V. CONCLUSION

In a clustering problem, it is always difficult to determine the number of clusters. This study shows that the auto clustering feature of SOM is more effective and objective than the K-means method. Thus, SOM can be utilized as the initial stage for web log analysis to determine the number of clusters and starting points. A combination of SOM with Simple K-means using Neural Network clustering method can also be used for web log analysis, since this combination has been proved to perform better than SOM and also better than simple K-means method in other applications. Thus the two-stage method, which first uses the SOM to determine the number of clusters and then K-means Neural Network algorithm to find the final solution of clustering, can become a robust clustering method for web log analysis.

## ACKNOWLEDGEMENT

We take chance to express our obligation and very thankful to all those who have helped us directly or indirectly to successful completion of this review paper.

## REFERENCES

1. Anand S. Lalani. “Data Mining of Web Access Logs”, Thesis Melbourne, Victoria, Australia, July, 2003
2. A. K. Jain, and R. C. Dubes, “Algorithms for Clustering Data”, Prentice Hall advanced reference series, Upper Saddle River: NJ, 1998.
3. R. Cooley, B. Mobasher, and J. Srivastava-,. “Data preparation for mining world wide web browsing patterns”, Knowledge and Information Systems Vol. 1, No. 1. Pages 5-32, 1999.
4. H. Ishikawa, M. Ohta, S. Yokoyama, J. Nakayama, and K. Katayama, “On the effectiveness of web usage mining for page recommendation and restructuring”, Web, Web- Services, and Database Systems.
5. E. Cohen, B. Krishnamurthy, and J. Rexford. “Improving end-to-end performance of the web using server volumes and proxy filters”, ACM SIGCOMM, pages 241-253, 1998.
6. R. Cooley, B. Mobasher, and J. Srivastava, “Data preparation for mining world wide web browsing patterns”, Knowledge and Information Systems, Vol. 1,No. 1, 1999.
7. World wide web committee web usage characterization activity. http://www.w3.org/WCA.
8. B.Amiri, M.Fathian. “Integration of self organization feature maps and honey bee mating optimization algorithm for market segmentation”, JATIT, Vol. 3, No. 3, Pages 70-86, July, 2007.
9. J.Srivastava. R.Cooleyz, M.Deshpande, Pang-Ning Tan. “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM, No.1, Pages.12-23, 2000.
10. J. Hartigan, and M. Wong, “A k-means clustering algorithm”, Applied Statistics, vol. 28, pp. 100-108, 1979.
11. K. A. Smith, and Alan Ng. “Web page clustering using a self-organizing map of user navigation patterns”. Decision Support Systems, 35(2):245– 256, 2003.
12. T. Kohonen, “Self-Organization and Associative Memory”, Springer-Verlag, New York, vol 10, Page 811-821, 1988.
13. K. M. Faraoun, A. Boukelif, “Neural networks learning improvement using the K-means clustering algorithm to detect network intrusions”, IJCI, Page 161-168, 2006.
14. Dehu Qi, Chung-Chih Li. “Self-Organizing Map based Web Pages Clustering using Web Logs” *Proceedings of 16th International Conference on Software Engineering and Data Engineering, SEDE,* Las Vegas, Nevada, Pages 265-270, July, 2007