# Recommendation of Web Pages for Online users using Web Log Data

**V.Chitraa[1], Antony Selvadoss Thanamani[2]**

[1](Assistant Professor, CMS College of Science and Commerce, Coimbatore, Tamilnadu, India
vchit2003@yahoo.co.in)

[2](Reader in Computer Science, NGM College (AUTONOMOUS), Pollachi, Coimbatore, Tamilnadu, India
selvdoss@yahoo.com

**Abstract:** ***World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. To reduce users browsing time lot of research is taken place. Web Usage Mining is a type of web mining which applies mining techniques in log data to extract the behaviour of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching and creating attractive web sites. Users' accesses are recorded in web logs. Because of the tremendous usage of web , the web log files are growing at a faster rate and the size is becoming huge. Web usage mining consists of three phases preprocessing, pattern discovery and pattern analysis. Soft Clustering is the most suitable method for web usage mining since same user can have more than one pattern and pattern analysis classifies the new user browsing in the knowledge base. Recommendations are given to the new user so that user's browsing time is utilized effectively. This paper describes the methodology for all phases of web usage mining.***

**Keywords-**Preprocessing, Fuzzy clustering, Session Identification, Recommendation, Web Log Mining

## 1. Introduction

The growth of internet is tremendous and abundant information is added daily which is readily available throughout the year and everywhere in this world. To attract more number of users the browsing to be made easier. Behaviours are discovered from web log files in which an electronic trail of data is left behind in the server whenever a user visits a web site and patterns discovered from which interested and purposeful web pages are to be made to the users.

Statistical tools and on-line analytical processing (OLAP) systems achieved limited success in understanding users previously that is until the concept of data mining was introduced. Data mining is the process of discovering hidden interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Web mining is defined as the discovery and analysis of useful information from the Web data. On-line businesses learn from the past, understand the present, and plan for the future by mining, analyzing, and transforming records into meaningful information.

Web Mining aims to discover useful information from web hyperlinks, page contents and usage logs. Based on the data web mining is categorized as web structure mining, web content mining and web usage mining. Web structures discovers knowledge from hyperlinks which represents structure of web site, web content mining discovers knowledge from web page contents and Web usage mining mines user access patterns from usage logs which record clicks made by every year. This helps to understand user's interest in a web site. The discovered patterns are usually represented as collections of pages, objects or resources that are frequently accessed by groups of users with common needs or interests. Some of the algorithms that are commonly used in Web Usage Mining are association rule generation, sequential pattern generation, and clustering.

The process of web usage mining can be separated into three distinct phases: pre-processing, pattern discovery, and pattern analysis. Web log data preprocessing takes the usage data recorded in server log as primary object. To increase the efficiency of mining preprocessing is to bedone and it takes 80% of the time in total process. It converts the Web log file into the data abstractions necessary for pattern discovery through extracting, decomposing, combining and deleting raw data. The general processes are data cleaning, users' identification, session's identification, path completion, transactions identification. The pattern discovery phase is applying data mining techniques on the preprocessed log data to discover some useful pattern. Some of the data mining techniques used are association, clustering, classification and so on. In this paper we focus on web users clustering and recommendations for personalizing the users. Methodical analyzing user's access information during a certain period of time will be able to understand the user's access mode and classification of web site visitors into different groups on the basis of their browsing pattern. In pattern analysis phase, the main aim is analyzing some of the mode, rule that have exhumed, to find out the patterns and rules we are interested.

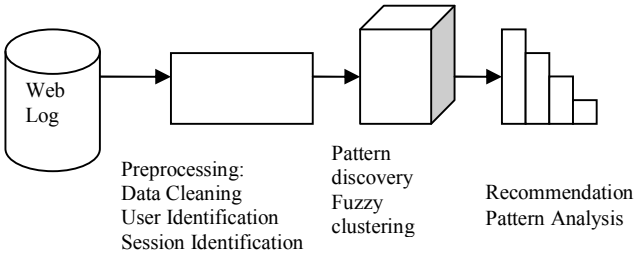


**Figure 1.** Phases of Web usage mining

This paper is divided into 5 sections and structured as follows. Section 2 analyzes works related to this paper. Section 3 describes the preprocessing process. Section 4 describes the clustering methodology to be implemented. Section 5 discusses how recommendations are given effectively to the users and section 6 gives conclusion and future enhancements.

## 2. Related Work

The focus of literature review is to study, compare and contrast the available preprocessing and clustering techniques. Preprocessing is used to increase the accuracy of mining results. To remove the invalid, auxiliary entries with image, robot navigation entries, records with unsuccessful status, data cleaning is done on the raw log file [7]. After cleaning the records users are identified from the cleaned log data. For websites with registration, username and password also entered in the log data. But due to privacy issues most of the users are reluctant to use those websites. So IPaddress and browser agent fields are used for finding the users who used the website. Site topology is also checked to identify a new user by the use of links. If the requested page is not reachable from any of the pages visited by the user then the user is identified as a new user in the same address [13].
Sessionization is the process of portioning sessions by following few heuristics. The simplest methods are time oriented in which one method based on total session time and the other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [5] to 24 hours [15] while 30 minutes is the default timeout by Cooley [13]. Another method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then the second entry is assumed as a new session. Third method based on navigation uses web topology in graph format. It considers webpages connectivity; however it is not necessary to have a hyperlink between two consecutive page requests. If a web page is not connected with previously visited page in a session, then it is considered as a different session. Different works were done by researchers for effective reconstruction of sessions. The referrer-based method and time-oriented heuristics method are combined to accomplish user session identification in [10]. A simple algorithm is devised by Baoyao Zhou [4]. Smart Miner is a new method devised by Murat Ali and team [12]. This framework is a part of their Web Analytics Software. Another effective method using Integer Programming was proposed by Robert F.Dell [14] in which all sessions is constructed simultaneously.

Clustering is the task of grouping together "similar" items in a data set. Clustering techniques attempt to look for similarities and differences within a data set and group similar rows into clusters. A good clustering method produces high quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high. Clustering algorithms could be classified into four main groups: partitioning algorithms, hierarchical algorithms, density-based algorithms, and grid-based algorithms [9].

**Partitioning Algorithms** attempt to break a data set of *N* objects into a set of *k* clusters such that the partition optimizes a given criterion. **Hierarchical Algorithms** create a hierarchical decomposition of a database and its decomposition can be represented by a dendrogram, which is a tree that iteratively splits the database into smaller subsets until each subset consists of only one object. Agglomerative hierarchical algorithms begin with all the data points as a separate cluster; followed by recursive steps of merging the two most similar (or least expensive) cluster pairs until the desired number of clusters is obtained or the distance between the two closest clusters is above certain threshold distance. Divisive hierarchical algorithms work by repeatedly partitioning a data set into "leaves" of clusters. **Density-based** algorithms locate clusters by constructing a density function that reflects the spatial distribution of the data points. Grid-based algorithms quantize the space into a finite number of cells and then do all operations on the quantized space. These approaches tend to have fast processing times, depending only on the number of cells in each dimension quantized in space, remaining independent of the number of data objects.

In [1] usage-based personalization using various data mining techniques have been discussed. A model-based clustering approach is discussed in [18] in which user's interest in a session is considered. The resulting clusters are used to recommend pages to the user. The paper uses different methods such as Poisson parameters and entropy to determine the recommendation scores. A usage based Web Personalization system called Web Personalizer using Web mining techniques to provide dynamic recommendations was proposed in [3]. Researchers have experimentally evaluated two different techniques such as PACT based on the clustering of user transactions and Association Rule Hypergraph Partitioning based on the clustering of clustering of page views for the discovery of usage profiles [2]. Formal Concept Analysis approach is used to discover user access patterns represented as association rules from web logs which can then be used for personalization and recommendation [19]. An improved Web page prediction accuracy by using a novel approach that involves integrating clustering, association rules and Markov models based on certain constraints has been presented in [8].

## 3. Preprocessing

Pattern Discovery can be done more accurately by using preprocessed data. Preprocessing is a complex process and takes 80% of total mining process. It is done to get more reliable data. There are 4 steps in preprocessing.

### 3.1. Data Cleaning

Raw web log data is noisy and irrelevant. So following log entries are to be removed in this stage

1. Entries like gif, JPEG are also downloaded along with user's request.
2. Unsuccessful status code log entries like <200 and >299.
3. Automated programs like web robots, spiders and crawlers.

### 3.2. UserIdentification

The log entry obtained after data cleaning is of the format {UIP, Date, Method, URL, Version, Status, Bytes, Referrer Url, BrowserOS, Timetaken }. This step is the most important step in preprocessing. Users are identified by using a method in which IP address and BrowserOperation Systemare used.

If two entries in web log consecutively are same, then they are considered as from the same user.

### 3.3 Session Identification

The sessions are used as data vectors in various classification, prediction, clustering into groups and other tasks. If URL in the referrer URL field in current record is not accessed previously or if referrer url field is empty then it is considered as a new user session. Reconstruction of accurate user sessions from server access logs is a challenge task and time oriented heuristics with a time limit of 30 minutes is followed. The set of user sessions are extracted as referrer based method and time oriented heuristics.
USS= {USID,(URI1,ReferrerURI1,Date1)…..(URIk, ReferrerURIk,Datek))}

Where USS is the user session set and $1 \leq k \leq n$, n denotes the amount of records in log set. Every record in log set must belong to a session and every record in log can belong to one user session only. After grouping the records into sessions the path completion step followsto find missing pages in the sessions due to 'back' key used by user and due to proxy server.

### 3.4 Computing the Time Taken

Reference length is the time taken by the user to view a particular page and it plays an important role. It is calculated by the difference between access time of a record and the next record. But this is not correct since the time includes data transfer rate over internet, launching time to play audio or video files on the web page and so on. The user's real browsing time is very difficult to analyze. The data transfer rate and size of page is also considered and the reference length is calculated as T =T'- bytes_sent / c

Where T' is the difference of access time between a record and the next one and bytes_sent is taken from log entry of a record and c is the data transfer rate. Transfer rate is calculated from the size and the upload speed of the server.

## 4. Pattern Discovery

Several reasons validate the idea of using pages visit duration as one of the weighting parameters. First, it reflects the relative importance of each page, because a user generally spend more time on a more useful page, because if a user is not interested in a page, he/she do not spend much time on viewing the page and usually jumps to another page quickly.
One important criterion to be considered in the choice of the clustering method is the possibility of creating overlapping clusters. This is a fundamental facet in Web personalization, where the ambiguity of the navigational data requires that a user may belong to more than one category or profile. Fuzzy clustering turns out to be a good candidate method to handle ambiguity in the data, since it enables the creation of overlapping clusters and introduces a degree of item-membership in each cluster [6].

### 4.1. Weight Calculation

In our weighting schema, both time length of a page and visiting frequency of a page are used to estimate its importance in a transaction, in order to capture the user's interest more precisely instead of binary which is typically used in other researches. This approach try to give more consideration to more useful pages, in order to better capturing the user's information need and recommend more useful pages to the user.

However, a quick jump might also occur due to the short length of a web page so the size of a page may affect the actual visiting time. Hence, it is more appropriate to accordingly normalize duration by the length of the web page, that is, the total bytes of the page. The formula of duration is given in Equation (1). Second, the rates of most human beings getting information from web pages should not differ greatly. If we assume a similarrate of acquiring information from pages for each user, the time a user spends on a page is proportional to the volume of information useful to him/her. As page duration can be calculated from web logs, it is a good choice for inferring user interest.

Frequency is the number of times that a page is accessed by different users. It seems natural to assume thatweb pages with a higher frequency are of stronger interest to users. A parameter that must be considered in the calculating the frequency of a page is the in-degree of that page (e.g. the number of incoming links to the page). It is obvious that a page with large in-degree has more probability to be visited by a user than a page with small one. Specially, in comparing two pages with same visiting rate, the page with small indegree is more interesting. The formula of frequency is given in Equation (2). We use time spent by a user for viewing a page and frequency of visiting as two very important pieces of information in measuring the user's interest on the page, so we assign a significant weight to each page in a transaction according to these definitions as Equation (3).

$$\text{Duration}(p) = \frac{\text{Total Duration}(p) / \text{size}(p)}{\text{Max (Total Duration)/Size(p)}} \quad (1)$$

$$\text{Frequency}(p) = \frac{\text{Number of visit}(p)}{\text{Number of visit}(Q)} * \frac{1}{\text{Indegree}(p)} \quad (2)$$

$$\text{Weight}(p) = \frac{2 * \text{Frequency}(p) * \text{Duration}(p)}{\text{Frequency}(p) + \text{Duration}(p)} \quad (3)$$

### 4.2. Matrix Transformation

Every user transaction is successfully transformed into a matrix with m rows and n columns and each cell value is the weight calculated for each user's interest, where *n i*s the total number of web pages visited in all users' transactions. $W_{ij}$ is the weight calculated for user's interest.

| | | | | |
|---|---|---|---|---|
| $W_{11}$ | $W_{12}$ | $W_{13}$ | --- | $W_{1n}$ |
| $W_{21}$ | $W_{22}$ | $W_{23}$ | ---- | $W_{2n}$ |
| $W_{31}$ | $W_{32}$ | $W_{33}$ | ---- | $W_{3n}$ |

### 4.3. Fuzzy MatrixTransformation

A web source matrix of the form R = ( Wij ) n x m can be converted into a web fuzzy matrix R' = ( rij ) n x n where every rij∈ [0,1]. This conversion is done by applying the transformation step [17]. That is by applying the below

formula on the web source matrix the web fuzzy matrix is obtained.

$$rij = 1 - c\sum_{k=1}^{m} |xik - xjk| (4)$$

The c in the formula is the factor that is selected to make the rij in the range of [0, 1] and it can be adjusted according to the practical situation.

### 4.4. Clustering

The next step is performing the clustering of the web users. The Clustering is based on the Relative Active Degree of the Web Users. For this clustering, the elements rij (i = 1… n; j = 1… m) that are different from each other in the web fuzzy matrix R. Using Euclidean distance similarity fuzzy matrix is created. When evaluating Clusters, put xi and xj into one cluster if rij is between 0 and 1. We can acquire the final matrix based on fuzzy equivalent relation and the final classification under different λ –threshold. If the fuzzy relation matrix is fuzzy equivalence relation, for any level of λ∈ [0, 1], the matrix is also equivalent relations.

## 5. Pattern Analysis

This is the online phase of the system. We can use the result to give a dynamic recommendations based on user's current visit. When a user visits the web site, the system match the pattern has been discovered, determine which category the user belongs to and then personalize the relational pages the category numbers interested to the user. The recommendation engine is the online component of a usage-based personalization system.The goal of personalization based on anonymous Web usage data is to compute a recommendation set for the current (active) user session, consisting of the objects (links, ads, text, products, etc.) that most closely match the current user profile. Recommendation set can represent a long/short term view of user's navigational history based on the capability to track users across visits.

During the online phase, when a new request arrives at the server, the URL requested and the session to which the user belongs are identified, the underlying knowledge base is updated, and a list of suggestions is appended to the requested page. A fixed-size sliding window is used as active session window to capture the current user's activities.In order to classify user session windows, the cluster that includes the larger number of pages in that session is considered. For this purpose Longest Common subsequences algorithm [11] is used to classify current user activates. According to the clustering algorithm discussed in offline phase there is a set of clusters as follows.

C = {C1, C2…..$C_N$}

Where $C_1$, C2 are clusters formed in pattern discovery phase and it consists of navigation patterns as

$C_1$ = {P1, P2….Pn}

Where P1, P2 are pages visited by similar users.For the active session window 'm' is the size fixed and A= {P1, P2…Pm} are the pages browsed by user currently. Compare A with clusters to find the suitable cluster by using LCS algorithm. For example, if a = {A,B,C,B,D,A,B} *and*b ={B,D,C,A,B,A}, Their LCS is g= {B, C, B, A}. If more than one cluster is matching based on LCS algorithm select a cluster that, if the difference between positions of last elements of longest commonsubsequence founded in the cluster and the position of first element of this sequence is minimized, the system must choices this cluster.

## 6. Conclusion

In this paper, we have made a systematic and complete methodology on offline and online phases of web usage mining. To acquire more accurate predictions preprocessing to be done on raw log data is discussed. User's interest is considered as a base for grouping similar users. Since soft clustering is the most appropriate one for usage mining to group users Fuzzy clustering is method and LCS algorithm for recommendations is discussed. The recommendations given by this methodology can help web site owners to provide personalized service to the users for their effective browsing. In future the methodology is to be implemented in a real data set.

## References


[1]. BamshadMobasher, Robert Cooley and JaideepSrivatsava, 2000. Automatic personalization based on Web usagmining. Commun. ACM, 43:142-151. DOI:10.1145/345124.345169

[2]. BamshadMobasher, Honghua Dai, Tao Luo, Miki Nakagawa, 2002.Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Commun. ACM, 6:61-82 DOI:10.1023/A:1013232803866

[3]. BamshadMobasher,H.Dai,T.Luo and M.Nakagawa, 2001. Improving the effectiveness of collaborative filtering on anonymous Web usage data. In Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization (ITWP01), Seattle.

[4]. Baoyao Zhou, Siu Cheung Hui and AlvisC.M.Fong, "An Effective Approach for Periodic Web Personalization", Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE, 2006.

[5]. Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web", Computer Networks and ISDN systems, 1995.

[6]. Castellano G.et al. "Mining user profiles from access data using fuzzy clustering" Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal, 2006

[7]. Chitraa. V, Antony SelvadossThanamani,"A Novel Technique for Sessions Identification in Web Usage Mining Processing", IJCA (0975-8887) Volume 34-No.9, November 2011.

[8]. Faten Khalil, Jiuyong Li and Hua Wang, 2008. IntegratingRecommendation Models for Improved Web Page Prediction Accuracy. Proceedings of the thirty-first Australasian conference on Computer science, 74:91-100

[9]. Houqun Yang, Jingsheng Lei, Fa Fu, "An approach of Multi-path Segmentation Clustering Based on Web Usage Mining", Fourth International Conference on Fuzzy Systems and Knowledge discovery, IEEE,2007.

[10]. Jose M. Domenech1 and Javier Lorenzo,"A Tool for Web Usage Mining", 8th International Conference on Intelligent Data Engineering and Automated Learning, 2007.

[11]. MehrdadJalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman , 2009, A Recommender System for

Online Personalization in the WUM Applications. Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II, San Francisco, USA pp741- 746

[12]. Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan " Discovering more accurate Frequent Web Usage Patterns", arXiv0804.1409v1, 2008.

[13].Robert.Cooley,BamshedMobasher, and Jaideep Srinivastava, " Web mining:Information and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[14]. Robert F.Dell, Pablo E.Roman, and Juan D.Velasquez, "Web User Session Reconstruction Using Integer Programming," IEEE/ACM International Conference on Web Intelligence and Intelligent Agent,2008.

[15]. SpilipoulouM.andMobasher B, Berendt B.,"A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis", INFORMS Journal on Comuting Spring, 2003

[16]. Sumathi C.P, R. PadmajaValli and T. Santhanam, "Automatic Recommendation of Web Pages in Web Usage Mining,"international Journal on Computer Science and Engineering Vol.02, No.09, 2010, 3046-3052

[17]. Sudhamathy .G., JothiVenkateswaran ," Matrix Based Fuzzy Clustering for Categorization of Web Users and Web Pages" International Journal of Computer Applications Volume 43– No.14, April 2012.

[18].ŞuleGündüz and M.T. Özsu,2003.A User Interest Model for Web Page Navigation. In Proceedings of International Workshop on Data Mining for Actionable Knowledge (DMAK, Seoul, Korea, pp 46-57.

[19]Vasumathi D. and A. Govardhan, 2005. Efficient Web Usage MiningBased On Formal Concept Analysis. IntelligentInformation Processing II IFIP International Federation for Information Processing, Volume 163/2005, 437-441.

## Authors Profile

**Mrs.V.Chitraa** is a doctoral student in Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu. She is working as an Assistant Professor in CMS college of Science and Commerce, Coimbatore. Her research interest lies in Database, Web Mining, Knowledge mining. She has presented many papers in conferences and published many papers in reputed international journals. She is an IEEE student member.

**Dr. Antony Selvadoss Thanamani** is working as Reader in NGM College, Pollachi with a teaching experience of about 25 years. His research interest includes knowledge management, web mining, networks, mobile computing, and telecommunication. He has guided 41 M.Phil scholars, attended 15 conferences, presented 35 papers, published about 8 books and many papers

.