ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# DOMAIN SPECIFIC APPROACH FOR WEBLOG MINING

**Kalyani Deotale*, Ms. Sonal Honale**
Affiliation: M.Tech CSE1; Assistant Professor

**DOI**: 10.5281/zenodo.47563

## ABSTRACT

Web mining is the application of data mining, chart technology, artificial intelligence and so on to the web data and identifies user's visiting behaviors and extracts their interests using patterns. It extract the behavior of users which is used in variety of applications like pre fetching, creating attractive web sites, personalized services, adaptive web sites, customer profiling, etc. The Common log formats or Extended Log Formats only records the visitors browsing activities rather than the details of the visitor's identity. A session is a sequence of activities made by one user during one visit to the site.There is a method in which is based on total session time and another method is based on single page stay time. The records with failed status code are also eradicated from logs. The changes in technology has made it possible to capture the users essence and interaction with the web. Due to large amount of incorrect information in the web log, the original log file cannot be directly used .

**KEYWORDS**: Preprocessing, Clustering, Web server logs, Web usage analysis, preprocessing, data cleaning, user identification.

## INTRODUCTION

In this internet era, web sites on the internet are useful source of information in everyday life. Therefore there is an enormous development of World Wide Web in its volume of traffic and the size and complexity of web sites. As per August 2010 Web Server survey by Net craft there are 213,458,815 active sites. Web mining is the application of data mining, chart technology, artificial intelligence and so on to the web data and identifies user's visiting behaviors and extracts their interests using patterns. The application of Web Usage Mining techniques in log data to extract the behavior of users which is used in variety of applications like pre fetching, creating attractive web sites, personalized services, adaptive web sites, customer profiling, etc.Web servers gathers data about user's interactions in log files whenever requests for resources are received. Log files records information such as client IP address, URL requested etc., in different formats such as Common Log format, Extended Common Log format which is issued by Apache and IIS.

Web usage mining includes three main steps: Data Preprocessing, Knowledge Extraction and analysis of extracted results. Preprocessing is an important step because of the complex nature of the Web architecture which takes80% in mining process. The raw data is pretreated to get reliable sessions for efficient mining. It includes the domain dependent tasks of data cleaning, user identification, session identification, and clustering and construction of transactions. Data cleaning is the task of removing irrelevant records that are not necessary for mining. User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. Path completion is used to fill missing page references in a session. Classifications of transactions are used to know the users interest ad navigational behavior. The second step in web usage mining is knowledge extraction in which data mining algorithms like association rule mining techniques, clustering, classification etc. are applied in preprocessed data. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge. Knowledge query mechanism such as SQL is the most common method of pattern analysis. Here the focus is on data cleaning and session identification process which is used to append lost pages and construction of transactions in preprocessing stage. The refined groups of pages are common user profiles we want. This theory is appropriate for clustering analysis because it provides an aggregation operator. It focuses on

data cleaning and session identification process which is used to append lost pages and construction of transactions in preprocessing stage. Web mining can be classified into three different types, which are Web content mining, Web structure mining and Web usage mining. Web content mining is the process of extracting and integration of useful data, information and knowledge from Web page contents. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site Web Usage Mining is a part of web mining which deals with the extraction of interesting knowledge from log files produced by web server. Web Usage Mining is also called Web Log Mining or Web Usage Analysis or Click Stream Analysis. This usage data provides the paths leading to accessed Web pages. This data is often gathered automatically into access logs by means of the Web server. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of the three phases the user can find the required usage patterns and use this information for the specific needs. The taxonomy of Web Mining is shown in Figure 1 below:

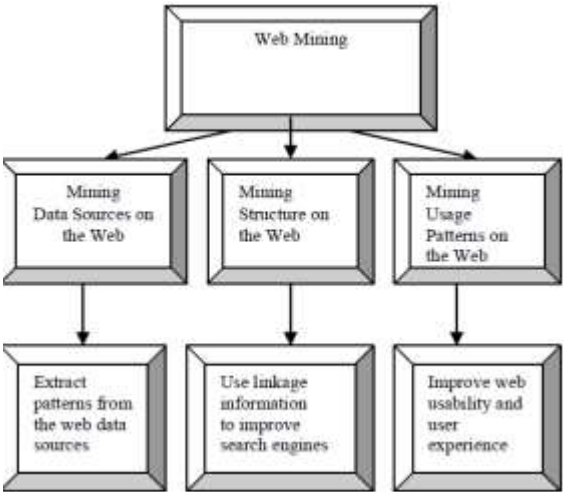


***Figure 1 :- Taxonomy of Web Mining***

## LITERATURE SURVEY

In the paper titled "**An Overview of Preprocessing on Web Log Data for Web Usage Analysis"** it says that Web has been growing as a dominant platform for retrieving information and discovering knowledge from web data. Web data is stored in web server log files. Web usage analysis or web usage mining or web log mining or click stream analysis is the process of extracting useful knowledge from web server logs, database logs, user queries, client side cookies and user profiles in order to analyze web users' behavior. Web usage analysis requires data abstraction for pattern discovery. This data abstraction can be achieved through data preprocessing. This paper presents different formats of web server log files and how web server log data is preprocesses for web usage analysis.

In the paper titled " **Data Mining for Web Personalization" it** present an overview of Web personalization process viewed as an application of data mining requiring support for all the phases of a typical data mining cycle. These phases include data collection and preprocessing, pattern discovery and evaluation, and finally applying the discovered knowledge in real-time to mediate between the user and the Web. This view of the personalization process provides added flexibility in leveraging multiple data sources and in effectively using the discovered models in an automatic personalization system. It provides a detailed discussion of a host of activities and techniques used at different stages of this cycle, including the preprocessing and integration of data from multiple sources, as well as pattern discovery techniques that are typically applied to this data. We consider a number of classes of data mining algorithms used particularly for Web personalization, including techniques based on clustering, association rule discovery, sequential pattern mining, Markov models, and probabilistic mixture and hidden (latent) variable models. Finally, we discuss hybrid data mining frameworks that leverage data from a variety of channels to provide more effective personalization solutions

In the paper titled "**Mining of Web Logs Using Preprocessing and Clustering"** it tells about The data pre-processing plays a major role in efficient mining process as Log data is normally noisy and indistinct. In data pre-processing method the rebuild of session and paths are going to complete by annexing lost pages. Additionally the transaction which explains the behaviour of users made accurate in pre-processing by calculating the time taken by the user to view particular page is accessed in the form of byte rate. By using web clustering various types of object can be clustered into different groups. The belief function similarity measures in algorithm include the clustering task by Dempster – Shafer's theory. The main aim of this work is to achieve pre-processing and clustering of web log and to improve the website performance.

In the paper titled "**Web Usage Mining:Discovery and Applications of Usage Patterns from Web Data**". It tells about web usage mining is the application of data mining tech-niques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. It describes each of these phases in detail. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the research and practice communities. This paper provides a detailed taxonomy of the work in this area, including research efforts as well as commercial offerings

In the paper titled **A Comparison of Document Clustering Techniques** we compare the two main approaches to document clustering, agglomerative hierarchical clustering and K-means.
Hierarchical clustering is limited because of its time complexity. K-means and its variants have a time complexity which is linear in the number of documents, but are thought to produce inferior clusters

## TECHNIQUE IMPLEMENTATION

Log data differs from other datasets used in data mining,. The main problem is to get a suitable dataset for mining..Web usage information takes the form of web server log files, or web logs. For each request from a user's browser to a web server, a response is generated automatically, called a web log file, log file, or web log.

This response takes the form of a simple single-line transaction record that is appended to an ASCII text file on the web server. This text file may be comma-delimited, space-delimited, or tab-delimited. A session is a sequence of activities made by one user during one visit to the site

### BASIC LOG FORMAT

**Remote Host Field**

This field consists of the Internet IP address of the remote host making the request, such as "141.243.1.172"
To obtain the domain name of the remote host rather than the IP address, the server must submit a request, using the Internet **D**omain **N**ame **S**ystem (DNS) to resolve (i.e., translate) the IP address into a host name.

**Date/Time Field**

The EPA web log uses the following specialized date/time field format: "[DD:HH:MM:SS]," where DD represents the day of the month and HH:MM:SS represents the 24-hour time, given in EDT

**HTTP Request Field**

The HTTP request field consists of the information that the client's browser has requested from the web server. The entire HTTP request field is contained within quotation marks.

**Status Code Field**

The status code field provides a three-digit response from the web server to the client's browser, indicating the status of the request, whether or not the request was a success, or if there was an error, which type of error occurred

.

**Transfer Volume (Bytes) Field**

The transfer volume field indicates the size of the file (web page, graphics file, etc.), in bytes, sent by the web server to the client's browser.

**Common Log Format**
The common log format (CLF or "clog") is supported by a variety of web server applications and includes the following seven fields:

Remote host field
Identification field
Authuser field
Date/time field
HTTP request
Status code field
Transfer volume field

**Identification Field**
This field is used to store identity information provided by the client only if the web server is performing an identity check

**Authuser Field**
This field is used to store the authenticated client user name, if it is required. The authuser field was designed to contain the authenticated user name information that a client needs to provide to gain access to directories that are password protected.

**3.2 EXTENDED COMMON LOG FORMAT**
The extended common log format (ECLF) is a variation of the common log format, formed by appending two additional fields onto the end of the record, the referrer field, and the user agent field.
**Referrer Field**
The referrer field lists the URL of the previous site visited by the client, which linked to the current page.
The referrer field contains important information for marketing purposes, since it can track how people found your site.

***User Agent Field***
The user agent field provides information about the client's browser, the browser version, and the client's operating system. This field can also contain information regarding bots, such as web crawlers. Web developers can use this information to block certain sections of the Web site from these web crawlers, in the interests of preserving bandwidth

Therefore the data should be pretreated and users' accessing behavior is to be constructed as transactions. These transactions are to be reliable. The Common log formats or Extended Log Formats only records the visitors browsing activities rather than the details of the visitor's identity. This means that different visitors sharing the same host cannot be differentiated. If there are proxy servers the problem became much severe. Users are identified easily by using Cookies or authentication mechanism. But users are not attracted by these types of sites due to privacy concerns.
If two records have varied IP address, then they are distinguished as two different users else if both IP address are same then User agent field is checked. If the browser and information of operating system's user agent field is different in two records then they are identified as different users. After users are identified the next step is identification of sessions. A session is a sequence of activities made by one user during one visit to the site. There are three heuristics available to identify sessions from users. Two heuristics are based on time and other is based on the navigation of users through the web pages.

***Time Oriented Heuristics*:** This method is based on total session time and the second is 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 [2] to 24 hours [10] while default time is 30 minutes. If the time exceeds 10 minutes the second entry is taken as a new session.

***Navigation Oriented Heuristics*:** This method uses web topology in graph format. It takes webpage connectivity, but it is not necessary to have hyperlink between consecutive page requests. Referrer based methods are used to append the missing pages. After session construction transactions are identified The transaction identification process depends on a split and merges process in order to look for a appropriate set of transactions that can be used in data mining.

**Density-based algorithms:** It starts by searching for core objects, and they are growing the clusters based on these cores and by looking for objects which are in a neighborhood within a radius of a given object. The advantage of these types of algorithms is that they can identify arbitrary form of clusters and it can filter out the noise. DBSCAN and OPTICS are density-based algorithms.

**Grid-based algorithms**: This algorithm uses a hierarchical grid structure to decompose the object space into finite number of cells. For every cell statistical information is accumulated about the objects and the clustering is achieved on these cells. The advantage of this approach is the fast processing time that is in commonly independent of the number of data objects. Grid-based algorithms are CLIQUE , STING , and Cluster

Fuzzy algorithms deduce that refusal hard clusters exist on the set of objects, but one object can also be assigned to more than one cluster. The best known fuzzy clustering algorithm is FCM .

## EXISTING SYSTEM

**Preprocessing**

Preprocessing of Web log data is a complex process . Log data is pretreated to get reliable data. The aim of data preprocessing is to select necessary features clean data by removing irrelevant records and finally transform raw data into sessions. There are four steps in preprocessing of log data

***A. Data Cleaning***

The data cleaning main process is removal of outliers or irrelevant data.. If a user requests a specific page from server entries like gif, JPEG, etc., are also downloaded which are not useful for further analysis are eradicated. The records with failed status code are also eradicated from logs. Removal process in our experiment includes

1. If the status code of each and every record when it is lesser than 200 and greater than 299 then those individual records are removed.

2. The cs-stem-url field is checked for its extension filename. If the filename has jpg, JPEG, CSS,gif and much more then they are removed.

3. The records which request robots.txt are removed

**B. Computing the Reference Length**

The time taken by the user to view a particular page is called as Reference Length . It is calculated by the dissimilarity between access time of a record and the next record. The user's actual browsing time is complex to analyze.

**C. User Identification**

The log file after cleaning is considered as Web Usage Log Set WULS = {UIP, Date, Method, URI, Version, Status, Bytes, ReferrerURL, BrowserOS}. Unique user identification is a complex step . The difficulty is due to the local cache and proxy servers. To overcome this cookies are used. The fields which are useful to find unique users and sessions are:-

- IP address
- User agent
- Referrer URL
-

Users and sessions are identified by using these fields as follows. If two records has same IP address check for browser information. If user agent value is same for both records then they are identified as from same user.

**D. Session Identification**

The goal of session identification is to partition the page accesses of each user into individual sessions. If the URL in referrer URL field in present 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 confronting task and time oriented heuristics with a time limit of 30 minutes is followed.

**Basic Probability Assignment For Each User**

Once data preprocessing, we discover that some sessions from identical user will overlap as a result of a user might perform identical task in several sessions. The probability measures however likely the user can perform the tasks known in the unique session.

Here probability is appointed to every unique session .It is the fraction of this unique session to the entire range of user sessions. The entire probability has measure one. This assignment is cheap since it captures the uncertainty among visits to distinct pages. The session by itself could be a semantically meaningful unit. It represents one or many tasks users tend to perform in one visit. Users typically have to be compelled to browse a group of pages, instead of a single page, to accomplish one task. Therefore, assigning a probability to a group of pages looks to suit perfectly the semantic which means of session.

## GREEDY CLUSTERING USING BELIEF

**Function**

A GCB algorithm for this clustering task using the similarity measure is defined in belief function. The greedy technique has been widely used in many algorithms as an efficient and effective way to approach a goal. In this process representatives of the clusters are done iteratively, so this present representative is totally separated from those that have been done in literature. An outline of the algorithm follows:

Input: K- number of clusters; S- a simple set of users,
Output: M- set of cluster representatives
begin
M = { *}
//select random users m1 into the common profile set
M = { m1}
For each user profile x _S – M, calculate the distance
between x and m1
Dist(x) = -ln (sim (x, m1))
For i=2 to K
begin
//choose representative m1 to be far from previous
representatives
Let m1ε S-M such that dist(m1) = max(dist(x)| x+ S-M
M = M*{mt}
// Update the similarity of each point to the closest
representatives
for each x ε S-M
dist(x) = min (dist(x), -ln(sim(x, m1)))
end
return M // M will contain a set of distinct
cluster representatives
end

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

## COMPARISON OF DIFFERENT CLUSTERING TECHNIQUES

***Table1: Factors According To Which Algorithms Are Compared***

| | Size Of Data Set | No. Of Clusters | Type Of Dataset | Type Of Software |
|---|---|---|---|---|
| ModifiedK-MEANS | Huge dataset and small dataset | Large number of clusters and small number of clusters | Ideal dataset and Random dataset | LNKnet Package and Cluster and Treeview Package |
| HC Algprithm | Huge dataset and small dataset | Large number ofclusters and small number of clusters | Ideal dataset and Random dataset | LNKnet Package and Cluster and Treeview Package |
| SOM Algorithm | Huge dataset and small dataset | Large number ofclusters and small number ofclusters | Ideal dataset and Random dataset | LNKnet Package and Cluster and Treeview Packag |
| EM ALGORITHM | Huge dataset and small dataset | Large number of clusters and small number of clusters | Ideal dataset and Random dataset | LNKnet Package and Cluster and Treeview Package |

***Table 2:- Relationship between number of clusters and the performance of algorithms***

| Number of clusters(K) | Performance | | | |
|---|---|---|---|---|
| | SOM | Modified K-means | EM | HCA |
| 8 | 59 | 61 | 62 | 65 |
| 16 | 67 | 71 | 69 | 74 |
| 32 | 78 | 84 | 84 | 87 |
| 64 | 85 | 89 | 89 | 92 |

***Table 3:- Relationship between number of clusters and the quality of algorithms***

| Number of clusters | Quality | | | |
|---|---|---|---|---|
| | SOM | Modified K-MEANS | EM | HCA |
| 8 | 1001 | 1118 | 1101 | 1090 |
| 16 | 920 | 1089 | 1076 | 960 |
| 32 | 830 | 910 | 898 | 850 |
| 64 | 750 | 840 | 820 | 760 |

***Table 4:- Affect of the data size on the algorithms***

| K=32 | | | | |
|---|---|---|---|---|
| DATA SIZE | SOM | Modified K-MEANS | EM | HCA |
| 36000 | 830 | 920 | 898 | 850 |
| 4000 | 89 | 95 | 93 | 91 |

***Table 5:-Affect of the data type on the algorithms***

| K=32 | | | | |
|---|---|---|---|---|
| Data type | SOM | Modified K-MEANS | EM | HCA |
| Random | 830 | 920 | 898 | 850 |
| Ideal | 798 | 810 | 808 | 829 |

## CONCLUSION

A data preprocessing system for web usage mining has been analyzed and implemented for log data. It has undergone various steps such as data cleaning, user identification, session identification and clustering. The quality of a website can be evaluated by analyzing user accesses of the website. To know the quality of a web site user accesses are to be evaluated by web usage mining

This preprocessing step is used to give a reliable input for data mining tasks. Perfect input can be created when the byte rate of each and every record is found. This algorithm lacks in scalability problem. Usage data collection on the Web is incremental. Therefore, there is a need for mining algorithms to be scalable.

## REFERENCES

[1] Bamshad Mobasher "Data Mining for Web Personalization,"LCNS, Springer-Verleg Berlin Heidelberg, 2007.
[2] Catlegde L. and Pitkow J., "Characterising browsing behavioursin the world wide Web," Computer Networks and ISDNsystems, 1995.
[3] Chungsheng Zhang and Liyan Zhuang , "New Path FillingMethod onData Preprocessing in Web Mining ," Computer and InformationScience Journal , August 2008.
[4] Cyrus Shahabi, Amir M.Zarkessh, Jafar Abidi and Vishal Shah"Knowledge discovery from users Web page navigation, "In.Workshop on Research Issues in Data Engineering, Birmingham,England,1997.
[5] Istvan K. Nagy and Csaba Gaspar-Papanek "User BehaviourAnalysis Based on Time Spenton Web Pages,"Web Mining Applications in E-commercce and E-Services, Studies in Computational Intelligence, 2009, Volume 172/2009, 117-136, DOI: 10.1007/978-3-540-88081-3_7 –Springer
[6] Jaideep Srivastave, Robert Cooley, Mukund Deshpande, Pang- Ning Tan"Web Usage Mining:Discovery and Applications of Usage Patterns fromWeb Data," SIGKDD Explorations. ACM SIGKDD,2000.
[7] Peter I. Hofgesang , "Methodology for Preprocessing and Evaluating the Time Spent on Web Pages," Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence2006
[8] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava,"DataPreparation for Mining World Wide Web Browsing Patterns," journalof knowledge and Information Systems,1999.
[9] Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, "Webmining:Information and Pattern Discovery on the World Wide Web,",InInternational conference on Tools with Artificial Intelligence, Newport Beach, IEEE,1997, pages 558-567.
[10] Spilipoulou M.and Mobasher B, Berendt B.,"A framework for theEvaluation of Session Reconstruction Heuristics in Web UsageAnalysis," INFORMS Journal on Computing Spring ,2003.
[11] Suresh R.M. and Padmajavalli .R. ,"An Overview of Data Preprocessing in Data and Web usage Mining ," IEEE, 2006
[12] Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path CompletionTechnique in Web Usage Mining,," International Symposium onComputer Science and Computational Technology, IEEE,2008.
[13] Yan Li and Boqin FENG "The Construction of Transactions for Web Usage Mining," International Conference on Computational Intelligence and Natural Computing, IEEE, 2009.
[14] http://news.netcraft.com/
[15] W. Wang, J. Yang, and R. Muntz, "Sting: A statisticalinformation grid approach to spatial data mining," 1997.
[16] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," pp. 94– 105, 1998.
[17] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases," in Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pp. 428– 439, 24–27 1998.
[18] C. Fraley and A. Raftery, "Mclust: Software for model-based cluster and discriminant analysis," 1999.
[19] J. C. Bezdeck, R. Ehrlich, and W. Full, "Fcm: Fuzzy c-means algorithm," Computers and Geoscience, vol. 10, no. 2-3, pp. 191–203, 1984.
[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in KDD, pp. 226–231, 1996.

[21] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data, (New York, NY, USA), pp. 49–60, ACM Press, 1999.