# An Effective Method to Preprocess the Data in Web Usage Mining

[1] **B.Uma Maheswari, [2]P.Sumathi**
[1] Doctoral student in Bharathiyar University, Coimbatore, Tamil Nadu, India
[2] Asst. Professor, Govt. Arts College, Coimbatore, Tamil Nadu, India
[1] umasharan7@gmail.com

## ABSTRACT

The Web mining field encompasses a wide array of issues, primarily aimed at deriving actionable knowledge from the Web, and includes researchers from information retrieval, database technologies, and artificial intelligence. Most data used for mining is collected from Web servers, clients, proxy servers, or server databases, all of which generate noisy data. Because Web mining is sensitive to noise, data cleaning methods are necessary. A data preprocessing system for web usage mining has been proposed in this paper. Data preprocessing includes data cleaning, user identification, session identification and path completion. The inexact data in web access log are mainly caused by local caching and proxy servers which are used to improve performance and minimize network traffic. The proposed method uses path completion algorithm to preprocess the data. We collect the datas from our college website and it is preprocessed based on the proposed method. The proposed path completion algorithm efficiently append the lost information and improves the consistency of access data for further web usage mining calculations.

**Keywords:** *Data preprocessing, Web usage mining, Path completion algorithm, Data cleaning, User session identification.*

## 1. INTRODUCTION

Web mining method is one of the effective method used in organizations to search the output from the large amount of the surface and WebPages from the hidden one. Many web mining algorithms are available to retrieve the WebPages. The main parts of web usage mining are: Data Preprocessing, Knowledge Extraction and analysis of results. Web Usage Mining consists of three main steps: data preprocessing, knowledge extraction, and results analysis. Raw data is highly susceptible to noise, missing values. The quality of data affects the data mining results. In order to improve the data quality, that the data is preprocessed. Usage of data preprocessing deals with the preparation and transformation of the data set. Cleaning, Integration, Transformation, Reduction are the methods involved in this. Data preprocessing has been studied extensively in the past decade (Cooley et al.,1999), and many commercial products such as Informatica S. Elo-Dean and M. Viveros,1997) and Data Joiner (Shahabi,1997) are applied in many areas.

In order to collect the data for preprocessing, much research has been done so far, e.g. the cookies (S. Elo-Dean and M. Viveros,1997) or the remote agent (Shahabi,1997) recognize the user session (Cooley et al.,1999) can do help to user identification, session identification and path completion. Data mining is one of the challenging to discover the large database. Data mining through these data preprocessing is increased and importance in industry. In competitive consumer markets, data mining faces the growing challenge of systematic knowledge discovery in large datasets to achieve operational, tactical and strategic competitive advantages. As a consequence, the support of corporate decision making through data mining has received increasing interest and importance in operational research and industry. As an example, direct marketing campaigns aiming to sell products by means of catalogues or mail offers(E.L. Nash,1992) are restricted to contacting a certain number of customers due to budget constraints.

The Web data is stored in Web servers, client machines, proxy servers or organizational databases. The primary data sources used in Web usage mining are the server log files which include Web server access logs, referrer logs and agent logs. Additional data sources that are also essential include the site files and meta-data, operational databases and domain knowledge. In some cases and for some users, additional data may be available in the client-side and proxy-server. Referrer logs contain information about the referring pages for each page reference. There are various types of Web data such as content data, structure data and usage data. Based on the type of data to be mined for analysis, Web mining can be further classified into Web content mining, Web structure mining and Web usage mining.

Data preprocessing is predominantly significant phase in Web usage mining due to the characteristics of Web data and its association to other related data collected from multiple sources. This phase is often the most time-consuming and computationally intensive step in Web usage mining. This process is critical to the success of Pattern discovery and Pattern Analysis. In short, the whole process deals with the conversion of raw Web server logs into a formatted user session file in order to perform Web usage mining(Sumathi et al.,2011).

The input for the Web Usage Mining process is a user session file, which is basically a pre-processed file and consists of information such as who accessed the website and what pages were accessed and for how long with their respective order. This user session file is first processed by removing outliers and irrelevant items from the raw server logs, identifying genuine and unique users from the server log and finally keeping the meaningful transactions within a user session file(P.Tan, and V. Kumar,2002).

The data cleaning process removes the data tracked in web logs that are useless or irrelevant for mining purposes. The request processed by auto search engines, such as Crawler, Spider, and Robot, and requests for graphical page content (e.g., jpg and gif images) are deleted because these image files are auto-downloaded with the requested pages. The user identification process analyzes the log file and clusters the users so that every user in the same group has the same access characteristics. Sessions identification Once the log files have been cleaned, the next step in the data preprocessing is the identification of the session. Session identification is the process of segmenting the user activity log of each user into groups of page references during one logical period called session.(Thanakorn Pamutha et al.,2012))

The paper can be organized as follows. Section II describes the related works , Section III describes the methodology used and section IV describes conclusion of the proposed work.

## 2. RELATED WORKS

Yan Li et al., (Yan Li Bo-Qin Feng and Yan Li,2002) presented a data preprocessing system for constructing the transactions in web usage mining. To implement transaction identification, the user sessions and the user access paths are extracted from the web access log and missing information is appended. These tasks are accomplished with the application of the referer-based method, which is an effective solution to the problems introduced by using proxy servers, local caching and firewall. Meanwhile, the reference length of accessed pages is calculated with the consideration of the time spent on data transfer over internet. Then two kinds of transactions are defined, i.e. travel-path transactions and content-only transactions. These two kinds of transactions are constructed by the maximal forward references (MFR) algorithm and the reference length (RL) algorithm, respectively.

Boqin Feng et al., (Yan Li et al.,2008) presented An implementation of data preprocessing system for web usage mining and the details of algorithm for path completion. After user session identification, the missing pages in user access paths are appended by using the referer-based method which is an effective solution to the problems introduced by using proxy servers and local caching. The reference length of pages in complete path is modified by considering the average reference length of auxiliary pages which is estimated in advance through the maximal forward references and the reference length algorithms.

Murat Ali Bayir et al., Murat Ali Bayir et al.,2006) introduced a new session reconstruction heuristic which is based on user web page requests logs. Smart-SRA has been experimentally shown to be better than previously developed reactive, time and navigation oriented heuristics. They did not allow page sequences with any unrelated (without any hyperlinks from the preceding page(s) to the next page) consecutive requests to be in the same session. Navigation oriented heuristics will insert artificial browser (back) requests into a session in order to guarantee that consecutive requests will have connectivity between each other. They also extend navigation oriented heuristics by using two time oriented heuristics. Another advantage of Smart-SRA is that it guarantees that all sessions generated will be maximal sequences and do not subsume any other session. They also implemented a novel agent simulator for generating simulated user sessions. They have compared the sessions reconstructed by Smart-SRA and previous heuristics against the simulated sessions generated by the agent simulator. They also defined a method to calculate the accuracy of the reconstructed sessions as a sequence –subsequence relationship.

Thanakorn Pamutha et al.,(Thanakorn Pamutha et al.,2012)) presented a brief introduction to WUM, apart from the data mining technologies and also the implementation of the preprocessing of web log files in NASA's web server. This study focuses on methods that can be used for the task of session identification from web log files. The work in this study also produces statistical information of user session. After preprocessing is completed, the result will be used for mining user access pattern, the future work involves various data transformation tasks that are likely to influence the quality of the discovered patterns resulting from the mining techniques like Association, Clustering, and classifications that may be applied only on to a group of sessions according to assumptions of users' intentions.

Chitraa et al., (V.Chitraa and Antony Selvdoss Davamani,2010) performed a survey on a selection of web usage methodologies in preprocessing proposed by research community. More concentration is done on preprocessing stages like session identification and path completion and have presented various works done by different researchers. Sumathi et al., (Sumathi et al.,2011) focused data preprocessing as a significant and prerequisite phase in Web mining. Various heuristics are employed in each step so as to remove irrelevant items and identify users and sessions along with the browsing information. The output of this phase results in the creation of a user session file. Nevertheless, the user session file may not exist in a suitable format as input data for mining tasks to be performed. They also focused on a design that can be adopted for preliminary formatting of a user session file so as to be suited for various mining tasks in the subsequent pattern discovery phase.

## 3. METHODOLOGY

Data Preprocessing plays a major role in Web usage mining process. Data preprocessing mainly depends on server log file.

The goal of preprocessing is to transform the raw click stream data into a set of user profiles(Demin Dong,2009). Data preprocessing performs a series of processing of web log file which includes data cleaning, user identification, session identification and path completion. The process involved in the data preprocessing is shown in the figure 1.

### 3.1 Data Cleaning

The data cleaning process involves removing the irrelevant data from the database log. This data may be in the form of requests from a non-analyzed source, data with missing attributes or the attributes that are not needed for the project goal. This step helps in reducing the size of the data to a great extent.

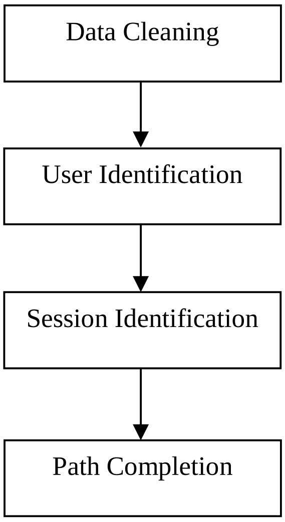

**Fig 1:** Steps involved in the data preprocessing

This reduction in size also helps in removing any false associations that could have been created because of this data. When a request to a web page is made, there are various attributes that are called and a lot of contents are loaded in that request. This includes the image files and graphics that are loaded with the web page because of the HTML tags. (Anand Sharma,2009)

Data Cleaning is a process of removing irrelevant items such as jpeg, gif files or sound files and references due to spider navigations. Improved data quality improves the analysis on it. The Http protocol requires a separate connection for every request from the web server. If a user request to view a particular page along with server log entries graphics and scripts are download in addition to the HTML file. Since the main objective of data preprocessing is to obtain only the usage data, file requests that the user did not explicitly request can be eliminated. This can be done by checking the suffix of the URL name. In addition to this, erroneous files can be removed by checking the status of the request (such as status code 404). Data cleaning also involves the removal of references resulting from spider navigations which can be done by maintaining a list of spiders or through heuristic identification of spiders and Web robots (P.Tan, and V. Kumar,2002). The cleaned log represents the user's accesses to the Web site.

The reference length of every access page plays an important role in the following processing procedures. Usually, the reference length of an accessed page is estimated by the access time of this page and the next one, i.e. the reference length of an accessed page equals the difference between the access time of the next and the present page. But with a more careful analysis, this difference includes not only user's browsing time, but also the time consumed by transferring the data over internet, launching the applications to play the audio or video files on the web page and so on. (Yan Li Bo-Qin Feng and Yan Li,2002) The user's real browsing time is difficult to be determined; it depends on the content of the page, the real-time network transfer rate, user's actions, computer's specifications and so on. If the time spent on data transfer is considered, the user's real browsing time of the accessed page is estimated by

$$RT = RT' - \frac{Bytes}{c} \qquad (1)$$

Where $RT'$ is the difference between the access time of the next and the present page, *Bytes* is the size of the present accessed page and *c* is the data transfer rate over internet. The algorithm used for data cleaning process is as follows(F.Yuan,2003):

```
        Algorithm Data Cleaning (Log File: Web log
file; Log File: Web log file)
  Begin
        While not eof (Log File) Do
        Log Record = Read (Log File)
        If ((Log Record.Cs-url-stem <>
gif,jpeg,jpgcss,js))

          AND (LogRecord.Cs-method= 'GET') AND
          (LogRecord.Sc-status = (200) AND
          (LogRecord.User-agent <> Crawler, Spider,
          Robot))

        Then Write (LogFile, LogRecord)
        End If
        End While
    End
```

By using this algorithm, data can be cleaned efficiently in the preprocessing process.

### 3.2 User Identification

User identification deals with associating page references with different users. To reduce network traffic and improve performance, the pages that are requested are cached by most Web browsers. Hence, when the user navigates backwards by using the "back" button, the repeat page access is not recorded in Web server log. Proxy servers provide an intermediary solution but the difficulty of user identification still persists. All requests coming from a proxy server have the same identifier even though the requests are put forth by multiple users. Two solutions for this problem are user registration data and use of cookies. One method to identify users is by means of the user id field in the server log files. The user registration data helps in capturing additional demographic information in addition to the data which is automatically collected in the server log. However, due to privacy reasons, many users prefer not to browse sites that require registration and logins. Sometimes user registration data is not compulsory and users may often provide incorrect information.(Sumathi et al.,2011)Hence, user identification becomes a complex task unless an exact user id is provided. In the absence of authentication mechanisms, the most well-known approach is the use of cookies.

The simplest method is to assign different user id to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. If the IP address of a user is same as previous entry and user agent is different then the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address . Caching problem can be rectified by assigning a short expiration time to HTML pages enforcing the browser to retrieve every page from the server.

### 3.3 Session Identification

A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. A transaction is defined as a subset of user session having homogenous pages. (V.Chitraa and Antony Selvdoss Davamani,2010)

The session identification splits all the pages accessed by the IP address which is a unique identity and a timeout; whereby the time between page requests exceeds a certain time limit. It is assumed that the user has started a new session. 30 minutes default timeout is considered. Algorithm used for session identification is explained below(F.Yuan,2003).

```
Algorithm Session den (LogFile: Web log file;
SessionFile: Session files)
  Begin
    SessionSet={}
    UserSet = {}
    k=0
    While not eof (LogFile) Do
        LogRecord=Read (LogFile)
     If (LogRecord. Time-taken>30min OR
        LogRecord.UserID not in UserSet)
     Then
        k = k+1
        Sk = LogRecord. Url
        SessionSet = SessionSet U {Sk}
        Write (SessionFile, SessionSet)
     End If
     End While
  End
```

A session is a sequence of page views by a single user during a single visit. A Session is the process of User activity record of each user in the log files. Session it shows single user visiting to web pages. In the ASP or ASP.Net session object is used, in this session object is used single user login status manipulation purpose. Same think should use in web usage mining to find how many sessions create a single user login to website. Session is partitioned after user identification. Session captures in two way

i. Time oriented
ii. Structure oriented

Time oriented is depends on the Time stamps or date and time of request in the server log file. In the time oriented session there are two types

i. The difference between First request and last request is < =30 minutes.
ii. The difference between First request and next request is <= 10. Using these two points we judge time oriented sessions.

Structure oriented capture in the referrer fields of the server logs. Structure oriented depends on Referrer fields is currently open or that user currently login referrer. It belonging to more than one “open” constructed session.

### 3.4 Path Completion

Path completion is necessary to be carried out due to the existence of local caching and proxy servers. The user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in user access paths should be appended. Path completion is depends on mostly URL and REFF fields in server log file. It is also graph model. Graph model represents some relation defined on Web pages (or web), and each tree of the graph represents a web site. Each node in the tree represents a web page (html document), and edges between trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site.

To append the missing pages, it is assumed that those pages are always accessed by using “Backward” button. The procedure of the path completion algorithm can be described as follows(Yan Li et al.,2008):

```
Last USID = 0; // The USID value of the previous record
```

```
Now ReferURI = “”; //The ReferURI of the current record
    L1: Getting the next record (i) in PS;
    if (Record(i).USID != LastUSID)
        {GOTO L3;}
    NowReferURI = Record(i).ReferURI;
    if (NowReferURI == Record(i-1).URI)
        {GOTO L3;}
    Getting Record(j); // j = i – 2

    L2: Record(j).RLength = ARLAP;
    Record(i-1).RLength=Record(i-1).RLength - ARLAP;
    Inserting Record(j) into PS’ according to the USID value;
    if (NowReferURI != Record(j).URI)
        {j - -;GOTO L2;}

    L3: Inserting Record(i) into PS’ according to the USID value;
    LastUSID = Record(i).USID;
    if (Record(i) is the last record in PS)
        {Outputting PS’;}
    else GOTO L1; //The End
```

This process makes certain, where the request came from and what all pages are involved in the path from the start till the end. The referrer plays an important role in determining the path for a particular request. The problem faced in this process is of the missing entries that mislead in tracking the request. But with the help of the referrer, the site topology and proper tracking of the web page requests, one can easily get the details of the path followed.

All of these processes of user identification, session identification and path completion together form the data-structuring phase of the classical data preprocessing scheme.

## 4. CONCLUSION

A data preprocessing system for web usage mining has been proposed in this paper. The process used in data preprocessing such as data cleaning, user identification, session identification, path completion. The algorithms used to preprocess the data has been analyzed. The proposed algorithms avoid the complicated procedure of mining site topology and don’t produce the user privacy issues. The modification of the reference length of the pages after path completion has also been implemented, it is very helpful for more accurate investigation on the user access pattern. The proposed method of data preprocessing system can prepare reliable transactions for the further web usage mining tasks. Thus our proposed method effectively preprocess the data in Web Usage Mining.

## REFERENCES

[1] Anand Sharma, “Determining Usage Patterns on RIT Web Data”, March 2008.

[2] C. Shahabi, A. Zarkesh, J. Adibi et al, “Knowledge discovery from users Web-page navigation”. In Workshop on Research Issues in Data Engineering, Birmingham, England, 1997.

[3] C. Shahabi, A.M. Zarkesh, J. Adibi, and V. Shah. Knowledge discovery from users web-page navigation. In Proceedings of 7th International Workshop on Research Issues in Data Engineering, Birmingham, England, 1997.

[4] C.P. Sumathi, R. Padmaja Valli , T. Santhanam, “An Overview Of Preprocessing Of Web Log Files For Web Usage Mining”, Journal of Theoretical and Applied Information Technology, vol. 34 ,no.2, December 2011. ISSN: 1992-8645.

[5] Demin Dong, “Exploration on Web Usage Mining and its Application”, IEEE,2009.

[6] E.L. Nash, The Direct Marketing Handbook, second ed.,McGraw-Hill, New York, 1992.

[7] F.Yuan, "Study on Data Preprocessing Algorithm in Web Log Mining," Proceeding of the Second International Conference on Machine Learning and Cybenetics, 2003.

[8] Marathe Dagadu Mitharam, “Preprocessing in Web Usage mining”, International Journal of Scientific & Engineering Research, Volume 3, Issue 2, February -2012 ISSN 2229-5518.

[9] Murat Ali Bayir, Ismail H. Toroslu, Ahmet Cosar, “A New Approach for Reactive Web Usage Data Processing”, Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006.

[10] P.Tan, and V. Kumar, "Discovery of Web Robot Sessions Based on Their Navigational Patterns”, Data Mining and Knowledge Discovery, 6:9-35, 2002.

[11] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. Knowl. Inf. Syst., 1(1):5–32, 1999.

[12] S. Elo-Dean and M. Viveros. Data mining the ibm official 1996 olympics web site. Technical report, IBM T.J. Watson Research Center, 1997.

[13] Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan, and Parinya Sanguansat, “Data Preprocessing on Web Server Log Files for Mining Users Access Patterns”, International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012, ISSN: 2046-6447.

[14] V.Chitraa and Antony Selvdoss Davamani, “A Survey on Preprocessing Methods for Web Usage Data”, (IJCSIS) International Journal of Computer Science and Information Security, vol. 7, no. 3, 2010.

[15] Yan Li Bo-Qin Feng and Yan Li, “The Construction of Transactions for Web Usage Mining”, International Conference on Computational Intelligence and Natural Computing, 2009.

[16] Yan Li, Boqin Feng, Qinjiao Mao, “Research on Path Completion Technique inWeb Usage Mining”, International Symposium on Computer Science and Computational Technology, 2008.