A Discovery on Web usage mining using Preprocessing

D.Durgadevi
Assistant Professor, Department of Computer Application,
Sri Akilandeswari Womens College,
Wandiwash, Thiruvannamalai. Tamilnadu, India.

Abstract:

Web mining is the technique of Data mining. Its applying data mining techniques to discover and collect the data from different web sites .Web usage mining (WUM) is automatic discovery and analysis of patterns in click stream and associated with data collected or generated as a result of user interactions with Web resources on one or more Web sites. A Data preprocessing is the process to converting the raw data into the data abstraction necessary for the further applying the data cleaning algorithm. For this paper ,I focus to the first phase in web usage mining .An overview of data preprocessing techniques aiming rectifying the irrelevant data and improve the accuracy of data presented in this review.

Keywords: *WUM-Web Usage Mining, raw data, Data cleaning Algorithm, web log files, data preprocessing*

I.INTRODUCTION

The World Wide Web rapidly grows at an far-fetched speed as information doorway. Web mining technologies are the suitable solutions for acquaintance discovery on the Web sites. Web mining is one of the important application of data mining techniques. It used to discover a patterns from the Web log data and to provide the desires of Web-based applications. Web Usage mining is the component of Web Mining. For typical data mining applications, the issues of data preprocessing plays a primary role in the entire mining process The outcome of Web Usage Mining can be used in personalized, improvement of system, website modification, business intelligence, usage categorization .The study going on data preprocessing of Web Usage Mining is a focus field nowadays. This paper attempts to present the process of data preprocess in the data preprocessing of Web Usage Mining. In this paper reviews the related researches in Web Usage Mining and examine the processes of data preprocessing in Web Usage Mining. There were loads of applications which utilize the web usage mining

for analyze the user navigation pattern. A data preprocessing is mandatory and important stage in web usage mining. The data cleaning and user identification are the important system in WUM Data -Preprocessing. The log files are data source to data preprocessing method. The main purpose of Data cleaning is to eliminating the inappropriate data. The main work of this processes are User identification that used to identify the accessing sites and displays what are all pages are accessed in web sites. The present study of data preprocessing technique is to be data cleaning and user identification. A variety of technique are provide for data cleaning but still there are issues remain in data collection and accuracy of user identification.

II.WUM PREPROCESSING

The raw web data were generally incomplete. The web log files were noisy, inconsistent and difficult to be used directly for pattern mining. The quality data gives quality output. The attributes of the quality data includes accuracy, completeness, consistency, accessibility, and timeliness. The WUM Preprocessing includes following techniques.

WUM Preprocessing Techniques:

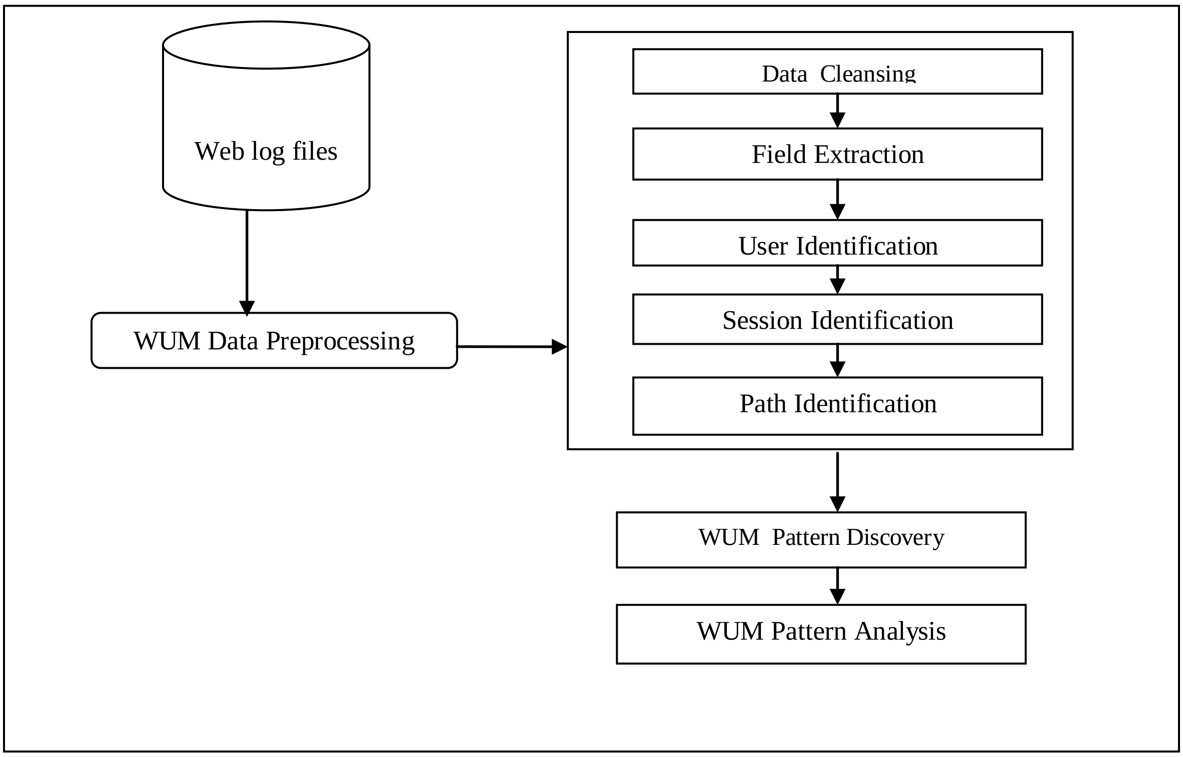
A .Data cleansing,
B .Field extraction,
C .User identification,
D .Session identification,
E .Path completion,

A .Data cleansing:

One of the most important steps in any data processing task is to validate the data values are correct or, at the very least, conform to some a set of rules. **Data cleansing**, **data cleaning** or **data scrubbing** is the process of detecting and correcting (or) removing the corrupt or inaccurate records from a record set.

B .Field Extraction:

The very important work that is to be necessary in all the preprocessing segment. It is obviously known as **field extraction**. The web logs file contain log entry which represents the single click stream. The log entry files may contains different fields which require to be separate out for the further dispensation The process of separating the different field from the single line of the web logs file is known as field extraction



C .User Identification:

This duty is significantly complicated by the survival of local caches, firewalls, and proxy servers. The data must be recorded by a Web server.These are not adequate for distinctive among dissimilar users and for distinctive among more than one visits of the same person.

- Common Log file Format (CLF)
- Extended CLF (ECLF)

D .Session Identification:

Here User session considered as the set of consecutive pages visited by a single user at a definite duration. For logs that extent long periods of time, it is measured as very likely that users will visit the Web site

more than twice. Its related to user identification, cookies and Session mechanisms both can be used in the

session identification, yet same problems also exit. The two criteria are usually considered in user sessions:

- superior limit of the session duration as a entire;
- superior limit on the time used up visiting a page.

If the above mentioned time exceeds that could be measured as a new session.

E .Path completion:

Consecutively to consistent identify unique user session, it should be determine if there are any important accesses that are not recorded in the web access log files. These are all the reasons in why issues occurring such substances are essential. Because the presence of web log Cache. If user clicks “back” to visit a page that has a copy, that will store in Cache log files, then browser will get the page directly from Cache files. In that page view will never be trail in access log, thus issues are the problem of deficient path which need renovation. The present methods are used to overcome this problem includes the cookies usage, busting of cache, and explicit registration of user. Cookies can be deleted by the user, cache busting defeats the speed of advantage that caching was created to grant and can be disabled, and registration of user is voluntary and users often provide failed information. If the referrer log is not clear, the site topology can be used to the same effect. If more than a page in user's history have a link to the requested page, it is implicit that the page neighboring to the previously requested page is the source of the new request.

III.RELATED WORK

Naga Lakshmi, Raja Sekhara Rao , Sai Satyanarayana Reddy et al., performed a work,” An Overview oreprocessing on Web Log Data for Web Usage Analysis”. In this paper web usage analysis have need of data abstraction for pattern discovery. This data abstraction could be achieve through data preprocessing technique. This paper presents various formats of web server log files and how the web server log data is preprocesses for web usage analysis.

Ankit R Kharwar, Chandni A Naik, Niyanta K Desai et al., performed a work, "A Complete Preprocessing Method for Web Usage Mining". The paper has numerous data preparation methods that preprocessing in order to identifying the unique users and users data session can be present to improve the performance features.

Mr. Sanjay Bapu Thakare et al., performed a work, "A Effective and Complete Preprocessing for Web Usage Mining". This paper describing the efficient and complete preprocessing of access stream previous to actual mining process be able to perform. The web log file collected from various sources undergo different preprocessing phase to make actionable data source. It will helps to automatic discovery of consequential pattern and relationships from access stream of user.

Li Chaofeng et al., performed a work," Research and Development of Data Preprocessing in Web Usage Mining".This paper presents the several data preparation techniques that can be accustomed to get better the performance of data preprocessing in order to identify unique users and user sessions. These methods and algorithms has been prove valid and efficient by experiment. Finally, this paper concludes the future research directions

Ke Yiping et al., performed a work, "A Survey on Preprocessing Techniques in Web Usage Mining". This survey fully focuses on the first stage of data preprocessing, which is important and should be perform prior to applying the data mining algorithms to sources of data. An overview of data pre-processing techniques aim at identifying the unique users, user sessions and transactions are all presented.

Marathe Dagadu Mitharam et al., performed a work," Preprocessing in Web Usage mining ".The main goal of this paper presented as the Analysis for interaction of user into various web .Web usage mining consists of following sections.

1) Pre-processing ,2) Pattern discovery ,3) Pattern Analysis .In this paper describes First phase in detail.

IV. CONCLUSION

WWW (World Wide Web) sites are plays most important role for web usage applications like Construction of site, Site adaptation, and site management and online Marketing . Also it is a most important tool for online advertisements . The quality of a website can be depends on the analyzing user accesses of the website. Log files are the best source to know user behavior. But the raw data contains unnecessary information like image access, false entries etc., which will affect the accuracy. Therefore, WUM preprocessing phase is an important work in web usage mining to make an efficient pattern analysis to get an accurate mining results.

V.REFERENCES

[1] Li Chaofeng. Data Source Analysis on Web Usage Mining. Journal of south-central university for nationalities, 2005(4):82-85(in Chinese)

[2] Kdnuggets. Software for web mining. http://www.kdnuggets.com/-software/web.html.

[3] Huaqiang Zhou, Hongxia Gao and Han Xiao "Rsearch on Improving method of Preprocessing in web log mining", IEEE 2010

[4] Tanasa D., Trousse B.. Advanced data preprocessing for intersites Web usage mining. IntelligentSystems, IEEE,2004(19): 59 – 65

[5] Ankit R Kharwar1, Chandni A Naik2, Niyanta K Desai, A Complete PreProcessing Methodfor Web Usage Mining (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

[6] Durgadevi D,Yamini G, A New Approach For Rectifying Erroneous Data In Web Usage Mining Using Preprocessing , IJSART - Volume 1 Issue 8 –AUGUST 2015 ISSN [ONLINE]: 2395-1052

[7] Sanjay Bapu Thakare et al. A Effective and Complete Preprocessing for Web Usage Mining (l1CSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010,848-851

[8] Rakhi Arya, Implementation of Intelligent Web Server Monitoring IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. III (Mar – Apr. 2015), PP 17-26