ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 7, July 2013

Accurate Analysis of Weblog Server File by Using Clustering Technique

Gurpreet Kaur

***Abstract*— Web usage mining is the process of data mining techniques to discover history for login user to web based application. Web usage mining is used to extract useful information from server log files. It is an automatic discovery of patterns in click streams and associated data collected or generated as a result of user interactions with one or more Web sites.**

***Goal* - Analysis for user interaction to various websites Web usage mining consists of different sections namely, Pre-processing, Pattern discovery and Pattern analysis. This research work describes pre-processing in detail. The present work would focus on extraction of the pattern of web log files of server to extract the pattern more accurately and extend the research from satisfactory to good condition by using the new fuzzy c means approach in backend database.**

***Keywords*— Data Mining data preprocessing: log file analysis, Fuzzy .**

I. INTRODUCTION

Web mining refers to use of data mining techniques to automatically retrieve, extract and analyse information for knowledge discovery from web log files and web documents[1].Then we collect all the information. It is called collection of data. After that we can go with the next level is preprocessing have explore the data from multiple log files,Data fusion, Data cleaning, Page view identification, User identification , Session identification, Path completion.

II. RELATED WORK

Web usage mining is a type of web mining, which exploits data mining techniques to extract valuable information from navigation behavior of World Wide Web users. The main task of data pre-processing is to prune noisy and irrelevant data, and to reduce data volume for the pattern discovery phase (Aye TT, 2011)[1].

Ting HI, Kimble C, Kudenko D(2007) et al.clarified that using two web usage mining techniques such as Automatic Pattern Discovery(APD) and Co-occurence pattern mining with distance measurement(CPMDM) for discover of potential browsing problems.[2].Chitra V and

DR. Devamani AS(2010) et al. Proposed that try to done the path completion, finding content, path set and travel path that is shows user interest[3].

Alam S, Dobbie G, Riddle P Observed that Describe a new web session clustering algorithm that uses particle swarm optimization. It will show that the algorithm performs better than the benchmark K – means clustering algorithm for new web session clustering.[4]

Ramya et al. (2011) proposed a complete pre-processing methodology having merging, data cleaning, user/session identification and data formatting and summarization activities to improve the quality of data by reducing the quantity of data. [5].

Mithram MD et all proposed that this research work describes pre-processing in detail. The present work would focus on extraction of the pattern of web log files of server to extract the pattern more accurately and extend the research from satisfactory to good condition by using the new fuzzy c means approach in backend database[6].

III. DATA PRE-PROCESSING

The pre-processing is the part of web usage Mining. We can also say that without pre-processing the web usage is not possible. The pre-processing is based on weblog files or server log files.

The data pre-processing is mainly used to remove the noise, transforming data and resolving any type of inconsistencies. If we want a meaningful pattern, then we need to use proper cleaning, transforming and structuring.

The pre-processing has following stages:

1)Data fusion and cleaning
2)User identification
3)Session data
   I. Time oriented
   II. Structure oriented
   III. Path completion

A. DATA CLEANING AND DATA FUSION

Data cleaning is used for remove the unnecessary and unwanted data and noise etc. Cleaning is the process which is

***Manuscript received July, 2013*.**
***Gurpreet Kaur**, Department of Computer Engineering, Yadawindra College of Engineering, Talwandi Sabo, Punjabi University, Patiala, Patiala, India.*

clean the unnecessary, unwanted data and improve the quality of the quantative data.
The data fusion have the multiple server are minimize the load of server. Multiple servers are marked to merge the log files from so many web and servers,etc[6].

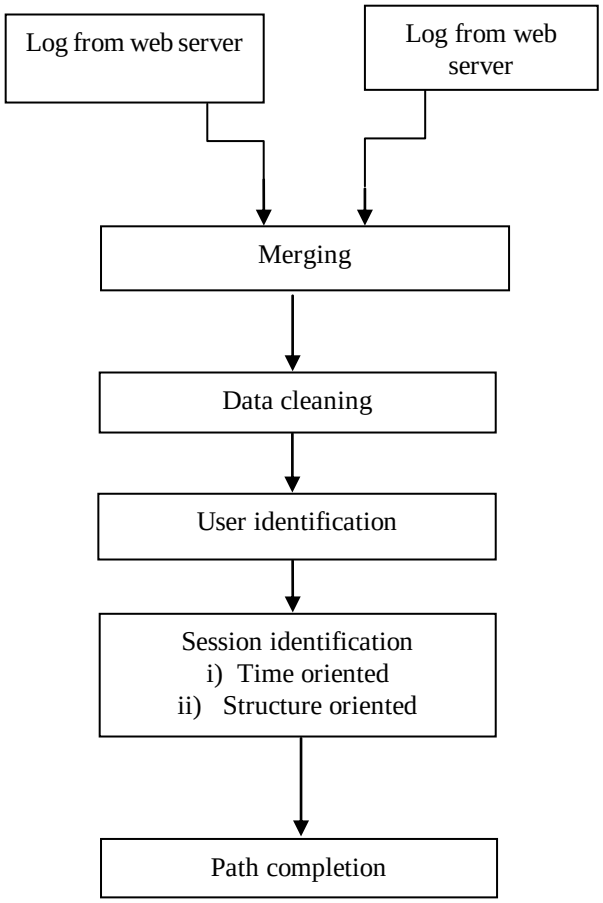

Figure (1) Diagram of pre-processing

B. *USER IDENTIFICATION*

User identification is possible through the ip addresses and agent. Ip address is shows who access the websites, pages and log files.if one ip address accessed more than one time it is not shows the same Ip address but we can find out the user through IP+AGENT address.

C. *SESSION IDENTIFICATION*

Every user have a time to access a log files, web pages. Every movement is their have record of a user when he/she entered to access a web page and when he/she left the web page.
The main motive of the session identification is to find out the every individual session of every user. There are two ways of session:

1. Time oriented
2. Structure oriented.

1.Time Oriented: it is depend on two ways:

a)The difference between the 1st entry and last entry is<=30mins.

b)The difference between 1st entry and 2nd entry is<=10mins.

2.Structure Oriented: a)consider a one session. b)Using the time oriented .we have carry two sessions because the 1st entry and the last entry having difference between>30mins

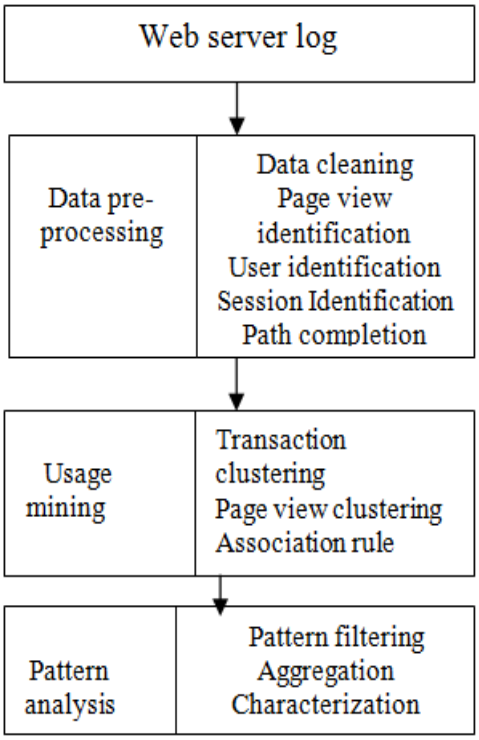

Figure (2) The process of web usage mining

D. PATH COMPLETION

It is confirming phase of pre-processing.in this phase we confirm that in which ways users visited the path. Path completion is used for reached and covering the user access path. If the path is not completed then it is caused of session identification. Path completion is based on REFF and URL in the server log files.
It is just like graph models. The graph having trees, the trees are like a websites. The nodes of tree are like web pages. The trees have edge which is like a websites.

AUTHOR NAME	PAPER NAME	TECHNIQUES/ ALGORITHM/ APPROACH	RESULT
Theint Theint Aye	Web log cleaning for mining of web usage pattern.	Field Extraction And Data Cleaning	Speed up extraction time when interested imformation is retrieved.
I-hsien Ting,Chris Kimble and Daniel Kudenko	Applying web usage ming techniques to discover potential browsing problems of user	Automatic Pattren Discovery (APD) and Co-occurrence pattern mining with distance measurement	Potential browsing problems of users can be discovered easily.
V.Chitraa ,Dr Antony Selvadoss Davamani	An Efficient Path Completion Technique For Web Log Mining	Path Completion Technique MFR and RL algo.	Content page set for analyzing user and so that modificatio n of sites can be done and Reliable input.
Shafiq Alam,Gillian Dobbie, Patricia Riddle	Particle Swarm Optimization Based Clustering Of Web Usage Data	New web session clustering algo. and particle swarm optimization	Performanc e of the algorithm is better than k-means clustering
Ramya C,Shreedhar a K S and Kavitha G	Preprocessing: A Prerequisite for Discovering Patterns In Web Usage Mining Process	Raw Data Of WUM Process	Reduce The Size Of Web Access Log Files Down To 73-82% and Offer Richer Logs Further Stages O WUM.
Marathe Dagadu Mitharam	Preprocessing In Web Usage Mining	Automatic Discovery Of Pattern and Associated Data Collected. Purpose Of WUM	After The Analysis Were Satisfactory And Contained Valuable.

IV. CONCLUSION

Preprocessing is important stage of WUM.In the present work, an attempt would be made to improve the quality of pattern of web log files. In the present approach under study, fuzzy c i.e., approach of clustering the web log files and removal of noise would be implemented.

REFERENCES

[1] Aye TT, "Web log cleaning for mining of web usage patterns". IEEE 490-494(2011)..

[2] Ting IH, Kimble C and Kudenko D, "applying web usage mining techniques to discover potential browsing problems of users".IEEE(2007).

[3] Chitra v and Dr SD Antony,"An Efficient Path Completion Technique For Web Log Mining"IEEE(2010).

[4] Alam S , Dobbie G and Riddle P, "Particle Swarm Optimization Based Clustering Of Web Usage Data" IEEE(2008).

[5] Ramya C, Shreedhara KS and Kavitha "Preprocessing: A prerequisite for discovering patterns In web usage mining process". International Conference on Communication and Electronics Information.V2,317- 321,IEEE(2011).

[6] Mitharam MD, "Preprocessing in Web Usage Mining" International Journal of Scientific & Engineering Research,Volume 3, Issue 2, February (2012).

BIBLIOGRAPHY:-

Gurpreet Kaur received her B.Tech degree in Computer Engineering from the College of Engineering, Rampura Phul (Bathinda) affiliated to Punjabi University , Patiala(Punjab) in 2011, and pursuing M.Tech degree in Computer Engineering from Yadawindra College of Engineering, Talwandi Sabo (Bathinda). Currently, she is doing her thesis work in Data Mining.