EasyChair Preprint

№ 1927

Survey on Pre-Processing Web Log Files in Web Usage Mining

Rita Roy and Apparao Giduturi

November 11, 2019

Survey on pre-processing web log files in web usage mining

Rita Roy Research Scholar, Professor G Appa Rao
Department of computer science and engineering
Gitam University, Visakhapatnam, AP

Abstract:

Web mining is to decide and extricate helpful data. In the web age web applications are expanding at gigantic speed and the web clients are expanding at exponential speed. As number of clients develops, site distributers are having expanding their data for pulling in and fulfilling clients. It is conceivable to follow the clients' quintessence and collaborations with web applications through web server log record and Web log document contains just (.txt) record. The information put away in the web log record comprise of enormous measure of disintegrated, inadequate, and pointless data. In view of enormous measure of superfluous information's accessible in the web log record, a unique log document can't be straightforwardly utilized in the web use mining. So pre-processing method is connected to improve the quality and effectiveness of a web log record. Various procedures are connected in pre-processing that is information cleaning, information combination, information joining. In this paper we will overview distinctive pre-processing system to recognize the issues in web log document and to improve web utilization digging pre-processing for example mining and investigation.

KEYWORDS: *data cleaning, data fusion, data integration, Pre-processing technique, Web usage mining, web log file.*

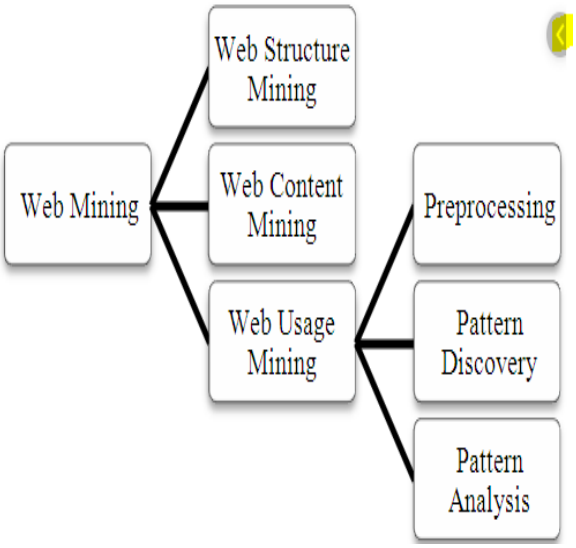
Introduction:

Amid the previous couple of years the World Wide Web has turned into the greatest and most well-known method for correspondence and data spread. It fills in as a stage for trading different sorts of data. The volume of data accessible on the web is expanding quickly with the unstable development of the World Wide Web and the approach of online business. While clients are given more data and administration choices, it has turned out to be progressively hard for them to locate the "right" or "intriguing" data, the issue usually known as data over-burden. Web mining is the use of Information mining methods to extricate learning from Web information including Web records, hyperlinks between archives, use logs of sites, and so forth. A typical scientific categorization of web mining defines three main research lines: content mining, structure mining and usage mining.

Web content mining is the procedure to find helpful data from the substance of a website page. Essentially, the Web substance comprises of a few kinds of information, for example, printed, picture, sound, video, metadata just as hyperlinks.

Web Structure Mining is the way toward deducing learning from the World Wide Web association and connections among references and referents in the Web. The structure of a run of the mill web chart comprises of website pages as hubs and hyperlinks as edges associating related pages. Web Structure mining is the way toward utilizing chart hypothesis to break down the hub and association structure of a site.

Web usage mining otherwise called web log mining is the utilization of information mining strategies on huge web log stores to find valuable. Earning about behavioural patterns and website usage statistics use insights that can be utilized for different web architecture undertakings. The principle wellspring of information for web utilization mining comprises of printed logs gathered by various web servers all around the globe.



There are four stages in web usage mining.

Data Collection*:* clients log information is gathered from different sources like server side, client side, proxy (intermediary) servers, etc.

Data Pre-processing: Plays out a progression of handling of web log record covering information cleaning, client identification, session ID, path completion and transaction identification.

Pattern discovery: Use of different information mining methods to handled information like measurable association, statistical analysis, pattern matching, clustering, etc.

Pattern analysis: when patterns were found from weblogs, uninteresting guidelines are sifted through. An investigation is finished utilizing the knowledge query system, for example, SQL or information blocks to perform OLAP tasks.

Regularly three primary information sources are utilized to gather log information for web use mining. Those are Server log, Proxy/intermediary server log, Client/ Browser log.

Server log: At the point when a web client demand a specific page on web, a section is signed into an uncommon record called server log document. This document isn't available by general web client, just authoritative individual or server proprietors can get to these records [1]. Server logs are considered as a most extravagant and solid wellspring of data to foresee client's behaviour yet it needs with numerous quality factors, for example, culmination and security issues.

Proxy/intermediary server log: A Proxy server is a server which goes about as a middle person between client's solicitations to other web servers. They are commonly utilized for storing data in to cache to improve route speed, managerial control, and security. Gathering intermediary level use information is comparative as gathering server level information.

Client/ Browser log: Web log information can likewise be gathered from client machine by incorporating java applets to the site, composing java contents or even altered browsers. Customer side logs are helpful to handle issues related to server logs like site page caching, session regenerating [2, 3].

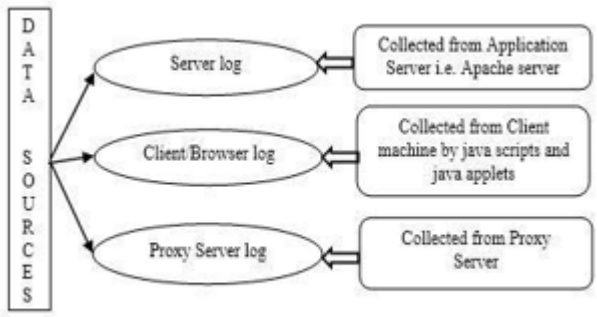


Figure 3: Types of log data and its sources

DATA PREPROCESSING

The fundamental strides of web utilization mining procedure are Data pre-processing, Pattern discovery and Pattern analysis [6]. Among them pre-processing is considered as an increasingly unpredictable and tedious procedure because of assorted nature of log information. It has been seen that pre-processing of log record takes additional time than different other stages of web utilization mining process [8]. It is important to perform pre-processing of log document to improve productivity and adaptability of essential information mining procedures connected on

log information. Web log information pre-processing should be possible in a few stages: Data Fusion, Data cleaning, Page see ID, User recognizable proof, Session Identification, Path completion, Transaction ID and Formatting [4, 6]. A few procedures like Data cleaning, User distinguishing proof, Session ID and Path completion have been talked about in detail one by one in following sub-segments.

Data cleaning

Web log document contains a lot of data into which some data are not applicable for web utilization mining reason so expulsion of these records is a fundamental advance [12].The status code beneath 200 or more 299 having passages ought to be evacuated [28]. The second step is the evacuation of graphical contents (sound, video, pictures) as they are downloaded with the mentioned page regardless of whether they are not expressly mentioned by the clients [3]. Graphical records are effectively recognized because of its document expansion (jpg, gif and so forth.). It is critical to expel these sort of solicitations since they are simply expanding the span of log document and nothing to do with investigation of client's navigational conduct. The third step is to clear log entries made by web robots (all over insinuated as web spider or web crawlers). Robots are an excellent kind of programming which is used by various web files to invigorate its arranged pages by getting to pages of a particular webpage in an incidental time interval [11]. Robot's revelation and ejection aren't as basic as graphical substance evacuation. A couple of methodologies for robot area are [16]:

i. Checking the client agent field where most robots announce themselves.
ii. Checking remote host name.
iii. Checking request of robot .txt file.

Client Identification

Client ID is one of the muddled errands because of the presence of neighbourhood/outside intermediary/proxy servers, cache frameworks, coordinate firewalls and shared web [7, 8]. There are a few techniques to recognize novel client is talked about underneath:

Client identification by IP address:

IP address is utilized to allot a one of a kind location to gadgets (PC, printers and so on.) taking an interest in system. IP address is signed into log document when a client hits a page. This location can be utilized to recognize various clients. In any case, if there should arise an occurrence of an intermediary/ proxy server when numerous clients demand a specific page at that point site server logged the same IP address (Proxy server IP) into the log document. For all intents and purposes, various clients are getting to that page. Storing likewise makes an issue to distinguish one of a kind client. At whatever point a client endeavour to get to recently gotten to a page, program show pages from the neighbourhood store and no passage is signed into the log document. A few strategies have been proposed to distinguish the right client and to address the above issues. Prabarskaite [18] unravel this issue by dismissing sections into log document on the off chance that they are from intermediary/proxy servers. As per him solicitations originating from intermediary/proxy servers can be recognized by domain name .for example in the event that solicitation is through the intermediary/proxy server the area definition contains "intermediary/proxy or cache" word. The issue with this methodology is that it misses some significant pattern from intermediary/proxy clients by dismissing their entrances in log document.

Client Identification by authentication information:

Enrolment information can likewise be utilized to recognize unique client. This is the most ideal approach to distinguish unique client by utilizing log name and rfname characteristics of the log record. If verification information (username and password) is asked while mentioning a page. This strategy isn't so much prominent on the grounds that client dependably attempt to stay away from such kind of sites [19].

Client Identification by cookies:

a cookie is a bit of information sent by a web server to customer machine when client demands a website page at the beginning time. This data put away in a content record as a text file on a customer machine with the browser. Cookies can contain helpful data with respect to a client so it is conceivable to effectively distinguish one of a kind client by utilizing it[9]. There is some situation where cookies don't work for example a few browsers do not bolster cookies, a few browsers impair cookies, cookies are erased by the clients and cookies are not logged by the web servers or erased by the servers.

Client Identification by user data:

For the most part, researchers utilize some heuristic strategies to distinguish interesting client. Mostly looking at agent field of Log document which contains working framework/operating system name and browsers name with respect to their version. On the off chance that two solicitations with the same IP/host addresses have distinctive browser's name or working framework/operating system at that point, there is a probability that postulations demand from two unique clients [7]. In spite of the fact that this strategy isn't dependable and results in perplexity for example in the event that a client is visiting two pages of the site by utilizing various browser at the same time on a single machine then this strategy will think about that two solicitations by various client even they are from single client.

Client Identification by site topology:

This strategy utilizes auxiliary webpage topology of the site to distinguish one of a kind client. Cooley et al. [7] has accepted that if a client demand a page that isn't open through its recently mentioned pages is considered as another client. This should be possible by utilizing the referrer trait of expanded log organization and connection data from site topology. Some circumstance where this methodology results in disarray for example on the off chance that client make a solicitation by utilizing bookmarked pages which are not associated through connections.

Session Identification

When client is distinguished there is have to recognize sessions. Session is set of solicitations done by single client for characterized term to a specific site. Fundamentally there are two different ways to discover sessions with respect to specific client [11]:

i) By utilizing verification data from clients, for example, cookies approaches or installed session id.
ii) By using heuristic methods.

These two techniques are likewise called as "proactive" and "receptive" strategies [20]. Proactive procedures make session dependent on session_id gathered from cookies. It makes a few cookies related issues which are recorded in client distinguishing proof techniques. In Reactive technique, sessions are developed from web log data [20]. The greater part of the analysts has taken a shot at the responsive techniques in light of the fact that proactive system depends on client's collaboration. Some receptive strategies are talked about underneath in detail.

Using through Time gap:

At the point when the time hole between two successive demands by a similar client is more noteworthy than certain edge then another session is made. Production of new session can be spoken to numerically by given condition [11]. Where and are time stamp of two back to back solicitation. The most famous limit esteem utilized by numerous specialists is 25.5minute. Anyway this can change from 10 minutes to 2 hour [3]. This esteem can be dictated by a few parameters like site topology, application type and so forth. The vast majority of the business sites and open source devices takes 30 minute 4 .limit esteem. Versatile or dynamic edge can likewise be utilized to improve productivity of session development [21, 22].

Using Through *Referrer* Attribute:

Session can be distinguished by referrer quality in expanded log design (Table 1). Assume x and y are two solicitations for continuous pages by same client and (a session), in the event that referrer of y was summoned already in that session S, at that point y is included session S generally another session is made with y as a first mentioned page [14].

Using through Time spent on observing page:

Cooley [7] arranges pages into two gatherings: Information pages and Navigational pages dependent on time spent on these pages. Data pages are those pages where clients are intrigued and Navigational pages are those pages which helps (for navigational reason just) to client's to reach at data pages. Clients invest more energy in educational pages than navigational pages. The term of time spend on navigational pages are littler. On the off chance that level of route page is accepted in log record, at that point most extreme length of route page is given by the equation: Where q signifies edge estimation of navigational pages, γ speaks to the level of navigational pages and μ indicates the mean estimation of watched term time for all pages in log document [7].

Path Completion:

There are odds of missing pages in the wake of building exchanges because of intermediary/proxy servers and caching issues [8] [9]. So missing pages are included as pursues: The page solicitation is checked whether it is straightforwardly connected to the last page or not. On the off chance that there is no connection with last page check the ongoing history. In the event that the log record is accessible in late history, at that point unmistakably "back" catch is utilized for reserving until the page has been come to. On the off chance that the referrer log isn't clear, the site topology can be utilized for a similar impact. On the off chance that numerous pages are connected to the mentioned page, the nearest page is the wellspring of new solicitation thus that page is added to the session.

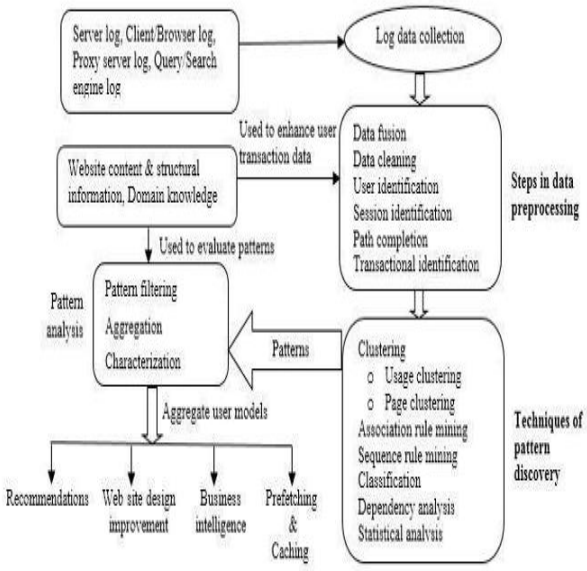


Figure 2: Web usage mining process and its applications [6, 11]

Literature survey and analysis:

In above areas different procedures utilized in pre-processing of log documents have been examined. Normally utilized sub-ventures of pre-processing are Data cleaning, client identification, Session identification and Path completion. Various scientists have acquainted different pre-processing systems with improve effectiveness and versatility of example revelation procedures. Some of them are talked about beneath:

Cooley et al. [7] have proposed strategies for information cleaning, client ID, session ID, and transaction identification. In spite of the fact that their strategies are adequate however a few heuristics are not suitable for complex sites.

Prabarskaite [15] proposed a superior cleaning technique. As indicated by him standard cleaning system isn't proper for edge pages containing sites. He connected two methodologies: propelled cleaning to improve web log mining and separating to evacuate superfluous connections. In this pre-processing procedure, creator did not play out some other strides of pre-processing like client identification session identification and so forth.

Tanana et al. [24] isolates pre-processing process in four stages: Data combination, Data

cleaning, Data structuration, and Data outline. Data combination creator joined various log records from various web servers and furthermore from website maps into a solitary log document. After that, they anonymised log document by encoding hostname. Further Data cleaning is performed by expelling demands for non-dissected asset, for example, sight and sound records (pictures, sound, video and so on.) and robot's created demands In Data structuration part creator have finished client recognizable proof by verification information or IP address, Session distinguishing proof by host and specialist, Page see ID by site map and so on. Finally, Data rundown step incorporates the design examination part by utilizing information speculation and accumulation. They didn't consider fruitless solicitation in information cleaning stage which is additionally required to evacuate to dispose of pointless estimations in later periods of weblog mining forms.

Castellano et al. [25] built up an instrument LODAP (Log Data Pre-processor) which takes log document as info and gives statistical analysis and client sessions as yield. This apparatus is partitioned into three modules: information cleaning module, information structuration module, and information filtering module. Information cleaning module evacuates interactive media records, status code, and robot's solicitation from log documents. In information structuration module clients are distinguished by validation information/IP address and sessions are recognized by time-based heuristics. The most extreme slipped by an incentive for session distinguishing proof has set to 30 minutes and minimum to 2 seconds between two back to back solicitations. Further in Data sifting most mentioned pages are held and least mentioned pages are dropped out dependent on limit esteem. Creators have finished practically every one of the means of information pre -processing with the exception of way consummation which is additionally a significant advance if there should arise an occurrence of reserve/intermediary server. They ought to likewise incorporate some more characteristics of log document with IP address for client recognizable proof successfully.

Robert et al. [26] presented another idea called whole number programming for better session recognizable proof. This strategy creates a session at the same time and delivered session better match to observational dissemination.

Yen li et al. [23] proposed a methodology for path completion by joining Maximal forward reference length and Reference length calculation. First Maximal forward reference is utilized to discover the arrangement of the page in client access way and it is additionally used to recognize the page, lastly Reference length calculation is utilized to discover whether the page is an instructive page or helper page. Finally by utilizing referrer field total way has been manufactured.

Kr suneetha et al[19] purely deal on weblog structure, data cleaning ,by encountering with the system error during each website visit .they performed the analysis upon it. In their work they did not approached for any algorithm on pattern discovery.

Jaideep Srivastava et al[20] in his research mentioned all the current existing work on web usage mining[wum] research project and product, mainly focused on web usage mining applications, personalization's, system improvement, site modification ,user characterization.

Xiang–ying li [30] has proposed a calculation named CSIA (Client and Session Identification calculation) for ID of client and sessions. This calculation incorporates exhaustive methodology by joining IP address, topology, browsers version, and referrer page to recognize interesting client with better precision and productivity. He proposed his calculation in JAVA language structure as it is useful for space usage. Anyway, this calculation is enduring with the abatement in working rate because of thought of numerous elements for distinguishing client. The rundown of writing audit is given in Table 2.

Author	Preprocessing techniques	Focused on	Remarks
Cooley et al. [7]	Data cleaning, user identification, session identification ,transaction identification	Transaction identification	Heuristic methods which are proposed not worked out in clumsy website.
Prabarskaite [15]	Advance data cleaning ,filtering and visulization	Data cleaning	Apart from this no other prepocessing steps used eg session identification ,user identification
Tanasa etal[24]	Data fusion, Data cleaning, Data structuration and Data summarization	Path completion not included	Wrong hhtp status code not removed efficiently
Castellano et al[25]	Data cleaning module ,data filtering module and data structural module	Path completion not included	All stages of preprocessing is covered.
Robert et al. [26]	Data cleaning and filtering, User identification, Session Identification	Session identification	By using integer programming they implemented a good sessions.
Yen li et al. [23]	data cleaning ,user identification ,session identification and path completion	Path completion	Maximal forward reference length and Reference length combined to find better path completion.
Xiang-ying li[30]	data cleaning ,client identification ,session identification and path completion	client identification ,session identification	Poor operating rate but having high efficiency and accuracy.
K r suneetha [19]	data cleaning ,client identification ,session identification and path completion	Web log structure ,data cleaning and user identification	No pattern discovery are focused here.
Jai deep Srivastava etal[20]	data cleaning ,client identification ,session identification and path completion	WUM applications ,personalization ,system improvement ,site modification user characterization	They give glance of webshift system to accomplish the web usage mining (WUM)from the server logs.

Reference

[1] Zidrina Pabarskaite, Aistis Raudys (2007), A process of knowledge discovery from web usage data: Systemization and critical review in Journal of Intelligent Information System, Springer Vol.28 Issue.1 Page(s): 79-104.

[2] C. Shahabi, F. Banaei-Kashani (2002), A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking in WEBKDD Third International Workshop on Mining Web Log Data, Page(s): 113-144.

[3] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos (2003), Web usage mining as a tool for personalization: A survey in User Modeling and User Adapted

Interaction journal, Vol. 13 Issues. 4 Page(s): 311-372.

[4] Cyrus Shahabi, Amir M.Zarkessh, Jafar Abidi andVishal Shah “Knowledge discovery from users Web page navigation, “, In.Workshop on Research Issues in Data Engineering, Birmingham, England,1997.

[5] Yan Li, Boqin FENG and Qinjiao MAO, “Research on Path Completion Technique in Web Usage Mining,,”, International Symposium on ComputerScience and Computational Technology, IEEE,2008.

[6] B. Naresh Kumar Reddy, M.H.Vasantha and Y.B.Nithin Kumar, “A Gracefully Degrading andEnergy-Efficient Fault Tolerant NoC Using Spare core”, 2016 IEEE Computer Society AnnualSymposium on VLSI, pp. 146-151, 2016.

[7] D. Tanasa, B. Trousse (2004), Advanced Data Preprocessing for Intersites Web Usage Mining in IEEE Intelligent Systems, Vol. 19 Issues. 2 Page(s): 59-65.

[8] Xiang-ying Li (2013), Data Preprocessing in Web Usage Mining in 19th International Conference on Industrial Engineering and Engineering Management Page(s): 257-266.

[9] Sanjay Bapu Thakare, Prof. Sangram Z Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", International Journalon Computer Science and Engineering, Vol. 02, No. 03, pp. 848-851,2010.

[10] Arvindkumar Dangi, Sunita Sangwan, " A new approach for user identification in web usage mining preprocessing", IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2287-8727, Vol. 11, Issue. 3, (May-June2013), Pages 57-61

[11] Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya (2013). Web Usage Mining: A Review on Process Methods and Techniques in Information Communication and Embedded Systems (ICICES), IEEE International Conference, Page(s): 40 – 46.

[12] R. Cooley, B. Mobasher, J. Srivastav (1999), Data preparation for mining world wide web browsing pattern in Journal of Knowledge and Data Engineering Workshop, IEEE, Vol.1 Page(s): 5-32.

[13] Zidrina Pabarskaite, Aistis Raudys (2007), A process of knowledge discovery from web usage data: Systemization and critical review in Journal of Intelligent Information System, Springer Vol.28 Issue.1 Page(s): 79-104.

[14] F. Facca, P. Lanzi (2005), Mining interesting knowledge from weblogs: a survey in Data and Knowledge Engineering, Vol. 53 Issue 3, Page(s): 225–241.

[15] Bing Liu,Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd ed. 2011

[16] Tasawar Hussain (2007), Hierarchical sessionization at preprocessing level of WUM based on swarm intelligence in 6th International Conference on Emerging Technologies, IEEE Page(s): 21-26.

[17] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos (2003), Web usage mining as a tool for personalization: A survey in User Modeling and User Adapted Interaction journal, Vol. 13 Issues. 4 Page(s): 311-372.

[18] Pabarskaite Z (2002), Implementing advanced cleaning and end-user interpretability technologies in web log mining in 24th International Conference on Information Technology Interfaces (ITI), Vol. 1 Page(s): 109-113.

[19] P.-N. Tan, V. Kumar (2000) Modeling of web robot navigational patterns, in: WEBKDD Web Mining for Ecommerce Challenges and Opportunities, Second International Workshop.

[20] Pabarskaite Z (2003), Decision trees for web log mining in Intelligent Data Analysis Journal, Vol. 7 Issue. 2 Page(s): 141–155.

[21] Renata Ivancsy, and Sandor Juhasz (2007), Analysis of Web User Identification Methods in World Academy of Science Engineering and Technology, Vol. 34, 2007.

[22] Berendt, B., Mobasher, B., Nakagawa, M., & Spiliopoulou, M. (2002). The impact of site structure and user environment on session reconstruction in web usage analysis in 4th WebKDD Workshop on Knowledge Discovery in Databases Edmonton.

[23] M. Chen, A.S. LaPaugh, J.P. Singh (2002), Predicting category accesses for a user in a structured information space in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Page(s): 65–72.

[24] J. Zhang, Ali A. Ghorbani (2004), The Reconstruction of user session from a server log using improved time oriented heuristic in 2nd Annual Conference on Communication Networks and Service Research IEEE, Page(s): 315-322.

[25] Yan LI (2008), Research on path completion technique in web usage mining in International Symposium on Computer Science and Computational Technology, IEEE, Vol. 1 Page(s): 554-559.

[26] D. Tanasa, B. Trousse (2004), Advanced Data Preprocessing for Intersites Web Usage Mining in IEEE Intelligent Systems, Vol. 19 Issues. 2 Page(s): 59-65.

[27] G. Castellano, A. Fanelli, M. Torsello, LODAP: A Log Data Preprocessor for Mining Web Browsing Patterns in Proceedings of the 6th Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Page(s): 12–17.

[28] R. F. Dell (2008),Web user session reconstruction using integer programming in International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, Vol. 1 Page(s): 385-388.

[29] Wahab, M. H. A., M. N. H. Mohd, et al. (2008), Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology.

[30] Xiang-ying Li (2013), Data Preprocessing in Web Usage Mining in 19th International Conference on Industrial Engineering and Engineering Management Page(s): 257-266.