ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

A Study on User Session Identification Techniques

Pavithra B[1], Dr. Niranjanamurthy M[2]

Assistant Professor, Department of MCA, Jain University, Bangalore, India [1]

Assistant Professor, Department of Computer Applications, MSRIT, Bangalore, India [2]

Abstract: In Web Usage Mining the log files of the web server plays a vital role because it stores the different users browsing patterns and this records becomes an important source of knowledge for discovering the user pattern. Web Usage Pattern is a process of retrieving the users browsing patterns by considering their page navigations. Mining techniques is applied to the user's behaviour for personalizing which is done based on transactions derived from user sessions. Sessionization is the process of identifying the user sessions, which is defined as set of pages visited by the same user within a given time of one particular visit of a web-site. This paper reviews the existing work done on the session identification techniques. An overview of available techniques for identifying the user sessions is being proposed for extraction of user patterns. By giving the overview of the techniques we can improve the quality of these techniques to be a novel one, by inventing new approaches and methods, and also we can work on by overcoming the flaws of the existing techniques which can be used as a highway for research and practice in this area.

Keywords: Include Sessionization, Heuristic, Web log data, Personalization, Smart Miner.

I. INTRODUCTION

Web Mining makes use of data mining techniques which automatically retrieves, extract and analyse the information from the web documents Web usage mining contributes a supportive system for designing the web site, providing personalized server and also helps the business persons to take a reasonable decision, etc. In order to better serve for the users, web mining applies the data mining, the artificial intelligence on the web data and traces users' navigational behaviour and then extracts the users' pattern. It is one of the most important areas in Computer Sciences because of its direct applications in e-commerce and Web information systems. There are three different domains that together club to give out the meaning of web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining.

Web content mining deals in retrieving the information from the content of documents or their descriptions. Web based technology, web indexing, resource construction and document text mining also falls under this.

Web structure mining deals with providing the conclusion on the information which is retrieved from the WWW and links between references and referents.

Web usage mining deals in extracting the interesting Patterns in web access logs.

Web Personalization is defined as the measure of taking the web experience of a particular user to personalise his interest by considering their preferences and observations, as and when they interact with the system and customising these contents for the benefit of the users.

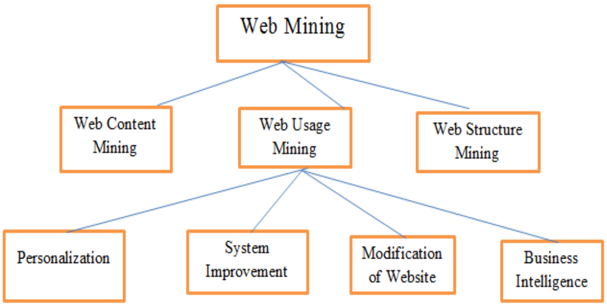

Fig.1. Web Mining categories

The internet world statistic says that there is a drastic increase in the number of web users as per the survey done in 2011. The result of the Netcraft survey says that the web sites are growing in a rapid speed which will be doubled every year. The web server log files records are also increasing because whenever the user access a web page an inlet will be created into those log files

There are four stages in mining those web server log files

- Data cleaning : As in the journal article [2] data collection is a process of collecting the log files which will be scattered in multiple servers
- Preprocessing : Log files contains noisy data which has to be eliminated by the process of data cleaning and

DOI 10.17148/IJARCCE

then followed by user identification, session identification, path completion and transaction identification is done as stated in the paper [8].
- Pattern Discovery: Most of the applications of data mining making use of pattern discovery for analysing statistical data, pattern matching, associating and clustering of knowledge.
- Pattern Analysis: Analysis of the required or interesting pattern is carried out based on the query mechanism.

Session identification is said to be considered as keeping track of all the pages visited by a particular user when the user opens a website in a particular duration. Based on the tracked data transactions of a specific user is constructed and it's said as subset of user session having homogeneous pages.

There will be so many hurdles in cleaning the server log files by removing the noisy data and reliably identify the user and his transaction session of the specific user. This paper presents a survey on the techniques that can be used in order to convert raw server logs into useful interesting sessions to meet the application of Web Mining.

II. RELATED WORKS

The focus of literature review is to study and compare the available Session Identification techniques. A user session identification is said to be considered as keeping track of all the pages visited by a particular user when the user opens a website in a particular duration. Personalization is defined as the measure of taking the web experience of a particular user to personalise his interest by considering their preferences and observations, as and when they interact with the system and customising these contents for the benefit of the users.

A user can have one or more than one sessions during the same duration. Once the user is determined then the pages he visited is classified and fed into logical clusters. This process of classifying into sessions is known as sessionization. There are many techniques available for session identification which is been discussed below.

- Jaideep Shrivastava et. al. [11] has discussed most of the research papers and the problem of user session identification.
- Reddy and his team as discussed in [12] have come up with a model for pre-processing the data, in which the model completely works on raw data cleaning, identifying the particular user and their session id's. Researchers has shown the problems on this model in maintaining the data quality and identifying the user and his session.
- Chintan R. Varnagar et. al. [14] has discussed most of the architecture of the system which takes either client or server side log data. In future an novel systems can be developed using this approach to give still more accurate and efficient result.
- Brijesh Bakaria et. al. [13], publishes a survey paper which gives the conclusion that no proper solution is available for session identification.
- Liu Kewen [15], proposed the algorithm for data cleaning and discusses the problem of user identification. It is not that easy to tie up the challenge of TB data.
- Mofreh Hogo et. el. [19] introduces the web usage mining of web users on educational web site, using the adapted Kohonen SOM based on rough set properties
- Sourabh Jain et. at. [20] presented paper is a review on data mining and fuzzy association rule to get the required data faster and efficiently.

A. A Novel Heuristic Technique

This technique makes use of heuristic algorithm which identifies the user of the website, user session and also the pages accessed by that user which will not be found in the relevant order in the server log. In this technique user session is identified by taking two things into consideration one is users browsing time and the other is his navigations between the pages. After identifying the session these sessions are applied with mining techniques for pattern discovery to know the interest of the user.

Experimental results of heuristic search is done by four phases [5]. As discussed in paper [5] the experimental result is given in four ways. The first step deals with the collection of data from the server; in this referred paper sample set of data is taken from charitable trust web server which has 9367 raw log records including noisy data.

The second step is cleaning of data to remove the irrelevant records which removes the burden of processing the user pattern, the records obtained after performing the second step is 1743.

The third step is user identification, in this step users are identified based on their IP address and their agent fields, by applying the clustering algorithm to the sample collected data 136 users are identified. The last step is session identification which is given using matrix as shown in figure 2.

The description of the figure is given as Users1, 2, 3 etc., indicates the user and their IP address. Session 1, 2 etc., gives the session created by using the heuristic method. As seen in the figure, for user 1 there are two sessions with the URL traversal 1-2 and 1-7. For user 2, only one session is created i.e., 5-6-7-8. For user 3, there is two sessions with URL traversals such as 2-4 and 7-8-9. For user 4, the session is 1 and URL traversal is 4-5-10. Then for user 5, the URL traversal is 2-6. User 6, the sessions are 4-6 and 9-10. For user 7, has 1-4-7 URL traversal.

DOI 10.17148/IJARCCE

Users	Sessions		URL							
		1	2	3	4	5	6	7	8	9
u1	s1	1	2	100						
u1	s2	2	0	0	0	0	0	2	100	
u2	s1	0	0	0	0	2	1	1	2	
u3	s1	0	1	0	2	100				
u3	s2	0	0	0	0	0	0	1	1	3
u4	s1	0	0	2	1	0	0	0	0	1
u5	s1	0	4	0	0	0	1			
u6	s1	0	0	0	3	0	5			
u6	s2	0	0	0	0	0	0	0	0	1
u7	s1	2	0	0	1	2	0	1	100	

Fig.2. Sample Matrix for Session Identification

B. Time And Navigation Oriented Sessionization

This technique involves three procedures in sessionization in which two procedure is been controlled by time and one by navigation. Time oriented procedure is simple one when compared to navigational based. In time oriented procedure few methods follows total session time and other few follows single page stay time. The authors Dr.Antony Selvadoss Thanamani and V.Chitraa in the paper [5] has taken 30 minutes as default timeout period. In the single page stay method time is calculated by taking the difference between the two time stamps of the user as default and the difference should be within 10 minutes if suppose the difference exceeds 10 minutes then the entry after that will be treated as a new session and the process will be repeated again to find out the time stamp difference.

The other procedure which is controlled by the user navigation uses web topology in graphical form. It takes pages connectivity into consideration, there is no compulsory that there should exists a hyperlink between two consecutive page requests. Cooley in the reference paper [1] has proceeded with a heuristic technique which takes URL references as a base and depends on user navigation, where the referred URL should be present in the same session of the specific user. As discussed by the authors J Guo, V Keselj & Q Gao in the paper [8] they have combined the heuristic time oriented methods and referrer based method to find the user session. An algorithm is been proposed in a simple way by Baoyao in the [7] which creates a session as a Paired URL and the time requested in the sequential form along with time stamps. This particular algorithm works when there are more URL's in a particular session of a single user. The author sticks on to a default time of 30 minutes for each session.

C. Smart Miner

Smart Miner is a frame work which is been implemented by the authors Jiawei Han and Micheline Kamber and discussed in brief in the reference paper [21]. As discussed in the paper the author and his team makes use of web analytics software as a part of implementation. The session created by SMART-SRA contains the information about the pages accessed from the server side. It is preceded in two systems first system follows Topology Rule and the other one follows Timestamp Ordering Rule. In the Timestamp ordering the stream of data is divided into shorter pages called as candidate sessions which is extracted by making use of stay time and session duration rule. In the second system the candidate session which is extracted from the first system is divided into maximum sub sessions, and the constrains of topology rule are followed and the backward browsing is been denied as it doesn't work on it. There is a chance of getting the hyperlink from the previous session then those pages are linked and added with the recently constructed session. In these cases the sessions are constructed sequentially one after other. The author and his team implement a simulator to simulate the web user which generates a specific website topology and also a user agent to retrieve from the client side and behaves as an original user. A unique character of the simulator is its ability to illustrate the user behaviours of a web agent. Time constraint is also considered as the difference between two consecutive pages is smaller than 10 minutes.

D. Integrated Programming for session identification

An another constructive method using Integrated Programming was proposed by the author Robert F Dell and his team, as in [9] all the sessions will be composed in parallel to identify the users and their sessions. We can also store the information such as user's id, IP address, URL and session time by using an simple algorithm.

This technique involves graphs for identifying the session and it gives out more accurate result, graphical representation is given as, the vertices refers the web pages and edges are depicted by the hyperlinks. Traversal of the graph is depicted as the web pages navigated by the user. The methods in Graph mining gives out the accurate result and also it's very simple to follow and implement. Another simple algorithm related to this technique has been proposed by Cristóbal Romero, Sebastián Ventura, Amelia Zafra, Paul De Bra [23] in which data cleaning and session identification are considered as a combined process.

III. ANALYSIS

A session is understood as a sequence of activities carried by a user when he is accessing a given site. To identify the sessions from the server log files which will be as a raw data including noisy data is a difficult task. This survey paper deals with different techniques available for Session identification based on time, navigations of the user's, heuristic techniques and algorithm, which is combined with accessing habit of web users and the characters of different pages.

Comparing all the above techniques for identifying the session, each of the techniques has some disadvantages

- In Heuristic Technique the proposed algorithm works fine only if the IP address is unique each time and it follows all the four phases if the user is not identified correctly then it leads to false session. This algorithm holds good only for limited number of users
- In Time based, the methods are not reliable because users may involve in some other activities after they open the web page and some circumstances like busy communication line, loading time of web page, web pages content size can't be considered and the default time is said to be an assumption time, so the calculation varies as the default time changes, the same users with a slight vary in default time leads to different sessions.
- In Navigational based method if a web page is not connected with previously visited page in a session, then it is considered as a different session and a new navigation is created
- A referrer based heuristics algorithm is been used the flaw of that algorithm is that the on the referred URL of a page should exists in the same session. If no referrer's is found then it is a first page of a new session. Again a new session is taken into consideration
- In Smart-Miner the backward browser moves are not taken into consideration and the pages without any referrers are determined in the candidate session from the web topology, then those pages are removed
- Time constraint is also considered as the difference between two consecutive pages is smaller than 10 minutes if suppose the difference exceeds 10 minutes then the algorithm gives a negative result. If hyperlink exists from the previously constructed session then those pages are appended to the previous sessions
- This method can handle the session one after the other, it cannot simultaneously handle the sessions if the user is accessing many websites at a time.

The novel and simple method for Sessionization can be calculated by taking the Average and Mean time to identify the user session. If the data will use the pre-processed than the identification of session will be accurate. The comparing experiment with traditional time based algorithm and its improvements shows that for better session identification, we should induce a novel algorithm that improves the accuracy of data pre-processing. The modified algorithm should be proposed on average time which completely depends on the web page, so we can say that using this time based algorithm with modifications will be used to separate the sessions, which is a better way to provide a constant values for sessionization.

A simplest algorithm can be implemented using clustered logic along with mining techniques to time based so that it adds the positive point for further implementation in an advanced way to give the accurate results

IV. CONCLUSION

A session is understood as a sequence of activities carried out by a particular user when he is navigating through a given site. To identify the sessions from the server log files which will be as a raw data including noisy data is a difficult task, because the server logs do not always contain all the information needed. Session identification is said to be considered as keeping track of all the pages visited by a particular user when the user opens a website in a particular duration. Based on the tracked data transactions of a specific user is constructed There are Web logs which do not contain enough information to construct the user sessions, in this case session identification becomes a complex task. The time-oriented or structure oriented heuristics can be used with a simplest algorithm using clustered logic along with mining techniques to time based so that it adds the positive point for further implementation in an advanced way to give the accurate results of session identification. Session identifications provides as one of the primary phases of pre-processing the data that later helps in user identification and pattern analysis to know the interest of the users and give out the interesting knowledge based on the user interest which is obtained by the process of pattern discovery and matching. Session Identification also provides secure commination channel for the user instead of betraying from the cookies. In certain situations passing the session id with the a URL parameters is said to be insecure so researchers should come up with the solution so as to make it more secured even id session id is passed as a URL parameter. This is also one of the challenging are for the researchers.By storing session identifiers in cookies, we can try to reduce the problem on sharing the link. This is conceptually called as session fixation, which involves intentional sharing. Session fixing means an attempt made to exploit the system which allows one person to find or set another users session id. Most of the

DOI 10.17148/IJARCCE

session fixation attacks are web based and they all depend on session identification being accepted from the URLS or post data.

This survey also help in identification of the end user behaviour with the help of the identification of the user IP address ,web link, user navigation, session usage, time and other techniques which the researchers can select best technique for identification of user behaviour more accurately for pre-processing data. We have presented the works done by different researchers so that a best technique can be taken up for further modification. Our research in future is to create more efficient session reconstructions and mining the sessions to give more accurate patterns for analysis of users.

REFERENCES

[1] S Bamshad Mobasher Robert Cooley, Jaideep Srivastava, “Automatic Personalization Based on web Usage Mining”, Communications of the ACM, New York, Volume 43,Issue 8, 20014

[2] Zahid Ansari, M.F. Azeem, Waseem Ahmed, A.Vinaya Babu, “Quantitative Evaluation of Performance and Validity indices for Clustering the Web Navigational Sessions”, proceedings of the IEEE /ACM International Conference on Data Mining & Knowledge Management Process, Vol. 3, No. 2, pp. 1-21, 2015.

[3] Dr.Antony Selvadoss Thanamani and V.Chitraa “A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing”, International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2013

[4] Alberto Cano, Amelia Zafra, Sebastián Ventura “Weighted data gravitation classification for standard and imbalanced data”, Cybernetics, IEEE Transactions on Cybernatics.2014

[5] SG Mathews,MA Gongora and AA Hopgood “Web useage mining with evolutionary extraction of temporal fuzzy association rules”, IEEE Transactions on Knowledge and Data Engineering, 17(6):734–749, 2013

[6] Abdul-Aziz and Rashid Al-Azmi, “Data, Text and Web Mining for Business Intelligence: A Survey”, proceedings of the IEEE /ACM International Conference on Data Mining & Knowledge Management Process, Vol. 3, No. 2, pp. 1-21, 2014.

[7] Baoyao Zhou, Siu Cheung Hui and Alvis C.M.Fong, “An Effective Approach for Periodic Web Personalization”, Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE, 2014.

[8] J Guo, V Keselj & Q Gao, “Integrating web content clustering into web log association rule mining”, In Proc. Springer, CCIS, Volume 3501, pp. 182-193, 2013.

[9] Robert F.Dell ,Pablo E.Roman, and Juan D.Velasquez, “Web User Session Reconstruction Using Integer Programming” , IEEE/ACM International Conference on Web Intelligence and Intelligent Agent,2008

[10] G. Arumugam and S. Suguna, 20015, "Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs", International Conference on Network and Service Security, IEEE, 1- 6.

[11] Jaideep Srivastava, Robert Cooleyz, Mukund Deshpande & Pang-Ning Tan, “Web Usage Mining Discovery and Applications of Usage Patterns from Web Data”, ACM-SIGKDD, Jan-2000.

[12] K. Sudheer Reddy, M. Kantha Reddy & V. Sitaramulu, “An Effective Data preprocessing Method for Web Usage Mining”, Feb-2013, IEEE

[13] Brijesh Bakariya, Krishna K. Mohbey and G.S. Thakur, “An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining”, Springer-2011.

[14] Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya & Jayesh N. Rathod, “Web Usgae Mining : A Review on Process, Methods and Techniques”, Feb-2013, IEEE

[15] Liu Kewen, “Analysis of Preprocessing Methods for Web Usage Data”, IEEE 2012

[16] Sheetal A. Raiyani, Shailendra Jain and Ashwin G. Raiyani, “Advanced Preprocessing using Distinct User Identification in web log usage data”, ISSN : 2278 – 1021, IJARCCE, Vol. 1, Issue 6, August 2012

[17] V. Sujatha and Punithavalli, “Improved User Navigation Pattern Prediction Technique From Web Log Data”, ELSEVIER-2012

[18] Hongzhou Sha, Tingwen Liub, Peng Qin, Yong Sun and Qingyun Liu, “EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining”, ELSEVIER-2013

[19] Mofreh Hogo, Miroslav Snorek & Pawan Lingras, “Temporal Web Usage Mining”, IEEE 2003.

[20] Sourabh Jain, Susheel Jain & Anurag Jain, “An Assessment of Fuzzy Temporal Association Rule Mining”, IJAIEM-2013

[21] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Second Edition, ELSEVIER Inc

[22] Weiss, S. M. et al. 2005. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer; 2014. ISBN 978-0-387- 95433-2.

[23] Ning Liu, Benyu Zhang, Jun Yan, Zheng Chen, Wenyin Liu, Fengshan Bai, Leefeng Chien. 2005. Text Representation: From Vector to Tensor. In: IEEE International Conference on Data Mining, ICDM, 2005. P.725-728.

[24] Cristóbal Romero, Sebastián Ventura, Amelia Zafra, Paul De Bra, “Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems”, Pergamon ,Journal Computers & Education pages 828-840,20012

[25] Jose M. Domenech1 and Javier Lorenzo, “A Tool for Web Usage Mining “, 8th International Conference on Intelligent Data Engineeringand Automated Learning ,2007

[26] Abdelhakim Herrouz, Chabane Khentout and Mahieddine Djoudi, “Overview of Web Content Mining Tools”, Proceedings of the fourth ACM International Conference on Web Search and Data mining, Pages 55-64, 2011

[27] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[28] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[29] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/ contrib. /supported/IEEEtran/

[30] FLEXChip Signal Processor (MC68175/D), Motorola, 1996.