ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Mining of Frequent Patterns using Session ID based on User Navigational Behavior

D.Gandhimathi[#1], Dr.N.Anbazhagan[*2]
[#] *Department of Computer Science and Engineering*
Alagappa University, Karaikudi, India
[1]Mathi.mathi65@gmail.com
[*] *Department of Mathematics*
Alagappa University, Karaikudi, India
[2] Anbazhagan_n@yahoo.co.in

***Abstract*— Web mining is one of the most capable areas in data mining; there are numerous areas of Web mining techniques, all can take about exciting fresh knowledge about information on the Web. Click stream analysis is one of the big tasks in the mining process but one avert cease is performed in this paper to predict the frequent results by the user interactions on the website. It has three important components such as first, each distinctive session is identified in the forth right manner before it goes to the log structure; Session Id is generated for every page-clicks and also identify the duplication. Second, every session related page clicks must be match up to the earlier set of the session page-clicks. Third, if its count value equals to the minimum support value it will be stored as a frequent page and as a final point in that frequent list, the permutations are achieved and it will be compared with every session. Here, diversion is realized from the pure click stream analysis by this work based on unique Session Id.**

Key words – web usage mining, click stream analysis, session identification.

I. INTRODUCTION

Web mining for the detection and the application of web usage patterns from Web data. Web mining is a data mining technique for match up to the contents of web pages to present more exact search engines. Web mining is one of the most capable areas in data mining, which means the Internet and World Wide Web are active sources of information. Web mining is the taking out of attractive and possibly useful patterns and implicit information from activity related to the web. Web mining comprise retrieving Web documents, selection and progressing of Web data, pattern discovery in sites, and analysis of the patterns. Web mining can be classified into three major concepts: web content mining, web structure mining, and web usage mining. Web content mining is the course of mining knowledge from the content of documents. Which basically includes the mining of web text oriented files; which is a form of resource detection based on the indexing of models. Web structure mining is the procedure of construing knowledge from the web links. Finally, Web usage mining also known as web-log mining, is the process of mining knowledgeable patterns from logs and other web usage information. Web usage mining is focused on web client behavior or modeling and forecasting how an online-user will use and gather information from the Web. Web usage mining has two different tracking such as general access pattern tracking and customized usage tracking. General access pattern tracking analyzes Web logs are used to understand access patterns by the user on the web. Customized usage tracking analyzes distinct trends; the purpose is to customize the websites to that users.

Click stream data is just tracking page hits. Every user moves from one page to another; that activity reveals the browsing behavior in the web. Click stream data captures data such as which page-links are clicked by users, how long a single user spends on each web page, and what types of searching terms are used by the user. In addition, if online-users complete a profile, more information can be gathered by matching these terms with searching manner. Finally, if the website belongs to a third party banner ad exchange, then the company might be able to track searching activity beyond its own website. This manner is referred to as click stream data. Click stream data allows without follow any searching behavior; which means users can search on the internet in arbitrary manner. On the Web, such random based analysis is very easier one. More and more companies are going to be turning to their click stream data and creating profiles based on customers' online purchasing data. The click stream analysis task can scan the entire log based structure and finally

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

provide the user satisfaction results with spends more and more time when the scanning within the large datasets. Click stream dataset which consist only numerical values that is series of data are available in the click stream log structure.

II. RELATED WORK

Zidrina Pabarskaite, Aistis Raudys recognizes a number of web log mining sub topics including specific ones such as data cleaning, user and session identification. Each sub topic is clarified, weaknesses and strong points are also conversed and possible solutions are presented in [11]. Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan present the current status of frequent pattern mining and discuss a little promising research directions; and believe that frequent pattern mining research has significantly widened the scope of data analysis and will have deep impact on data mining techniques and applications in the long run in [3]. Zdravko Markov and Daniel T. Larose present a composed of four chapters covering the different steps of the data mining process: log file content and format, log file preprocessing, exploratory data analysis, and finally, modeling for web usage in [10]. DeMin Dong presents the approach based on data cube stresses on turning web logs into structuralized data cube which can understand multi-angle comprehensive mining and analysis as well as introduce a diversity of mature data mining technologies in [1]. Robert Cooley, Bamshad Mobasher, J aideep Srivastava presents several data preparation techniques in order pattern instead of confidence in [2]. Theint Theint Aye mainly focuses on data preprocessing stage of the first phase of web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file in [7]. Ran Tang and Ying Zou present an approach that routinely identifies service composition patterns from a variety of applications using execution logs in [5]. Tomás Arce, Pablo E. Román, Juan Velásquez, Víctor Parada Presents a heuristic approach based on simulated annealing for the sessionization problem. Using this approach, it has been possible to decrease the processing time up to 166 times compared to the time that is required for the integer programming model in [8].

III. PROPOSED WORK

The click stream analysis, all user-clicks are stored within the log format structure. Click streams are online-users' paths through a web site. Click stream analysis illustrates how a web site is navigated and used properly by its users. So it must progresses the overall log based structures then performs any grouping based process and then gets out the results. Click stream data can be considered as a rich source of information, because it contains navigational information of the visitor on the website. The clearest reason for investigating click streams is to mine important information about what people are doing on particular site. Examining individual click streams will present the information and need to make content related decisions. However it is difficult to analyze since it is available as unstructured data and many different formats depending on the web server. Weblogs can be mined as .txt files from the server. For example, analyzing only the access logs then all other types of logs like error logs; so more time utility is needed here to do the analyze of all surfing behavior of the online visitor. So the forth right manner of task is performed in this paper to get rid the click frequent results. But in the proposed methodology has a unique structure when compared with the existing techniques; that unique structure is based on three components such as first, every unique session is realized in the forth right manner with divergent Session-Id (*SI*) then the session based results will be stored in the log structure. Second, every session elements is compared with the previous collection of the session elements which has values; if which are available then the values are incremented otherwise not incremented. Third, when its count value reaches the minimum support count value it will be stored as a page of frequent list and finally in that frequent list, the combinations are performed then it will be compared with every individual session.

This means if the same combination is checked in whatever sessions are available. Here, consider the sample set of online-page clicks below:

TABLE 1: THE SAMPLE LIST OF ONLINE-PAGE CLICKS

Sessions	Requested-Pages
Session 1	rp1,rp2,rp3
Session 2	rp2,rp4
Session 3	rp1,rp2,rp4
Session 4	rp3,rp5
Session 5	rp4,rp5
Session 6	rp1,rp3,rp4,rp5
Session 7	rp1,rp4

Assume seven diverse of sessions for getting the count of every individual page-click. From the above sample table 1, this represents every unique session and its own number of online-page clicks. Here, the first row of the table 1 reveals first session values which means page click; if that page is present the value is initialized by 1 otherwise, value is null. From the second row wherever the request-page is present it will be changed as a

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

incremented value otherwise all are null; in the second session only two page clicks are presented {rp2:2,rp4:1} here one important task is running that is the first session has {rp1:1,rp2:1,rp3:1} and the remaining are like {rp4:0,rp5:0}. Session 2 has {rp2:2,rp4:1}, here rp2 has been presented in the previous session also; then consider rp4, but the same page click is not available in the previous session because it has value zero that means null. Consider next session {rp1,rp2,rp4} but one important thing is, how can decide the value of rp1 because the page-click is not present in the previous session so check whether the before of previous click is present or not. This complexity is not in the rp2 page click because it has the previous visit and also rp4. For example the set {rp1:1,rp2:1,rp4:1} in the previous collection then it will be changed like {rp1:2,rp2:2,rp3:1,rp4:1} after the next session entrance. The same above procedure is followed at the end of the process.

A. *Session identification*

First of all should identify the every unique session data based forth right manner before it reaches to the log structure. Every session is identified before the browser is closed; after the closing of browser by the individual user, if the same user watches the pages again with some other new pages of that website then it will be considered as a new session in this analysis. In this proposed work, unique *SI* is created for every distinct page-click; if the same page is clicked by the same user again in that particular same session then it will not be considered as a count because that page click has its individual unique session id and that has been registered already. By this type of manner the duplication omitted involuntarily that means the previously clicked page is not come again in further on that identical session. Running set has all those clicked pages which means every clicked page is calculated on the running set. After that session is called previous session set and the next session set is called as running set. Likewise every unique session is identified based on the forth right manner before all reaches its log structural format.

B. *Comparisons of the Previous Set with the Running Set of Pages*

Click stream data can be obtained from page clicks directly which referred to as hits and their associated information such as, user agent, IP address, URL, status, timestamp, number of transferred bytes, referrer, cookie data recorded in web server log files. These log details of web site visitors can be stored as particular file or can be divided as access logs, error logs, distinct logs and more. Site administrators usually have complete control over these files. Here, every running set has number of page clicks with unique session-id (*SI*); this session id is created automatically when the online visitor clicks the page once. Here, the running set is compared with the previous session set if the previous set has {rp1:3},0,{rp3:3},{rp4:4},{rp5:3} so this session only contains pages 1 and 3 through 5. The comparison is running here with the running session set like {rp1:4},0,0,{rp4:5},0; from this result page 1 and 5 only incremented and all the other clicks are null so there is no hitting on those pages by the online visitor. Every visitor surfing the website with arbitrary manner so which are not a predictable one by the specialist and other people. Whenever, the hit is made on pages that will be gathered as a collection of pages for individual session.

C. *Storage of Results*

Based on the consideration of the support value, the results will be stored on the set. Here, the support value (s) is considerably 3 which is denoted by symbolically s=3; when the frequency count will reach to the support value that should be considered as the frequent page.

$$\text{Frequency Count} \geq \text{support value}(s) \qquad (1)$$

The condition (1) is processed to check the above statement if the result is true, that will be added to the frequent list otherwise it will not be considered; here one important matter is handled by this work that is whatever page-clicks are not there, then all are omitted by this work.

IV.CONCLUSIONS

More web usage mining technologies are available to provide the frequency list of records from user's navigational manner; here a new Session-Id (*SI*) based work is done to produce the frequency based results without surfing the log file structure as whole. This type of identification is different with the click stream based work because it is the forth right manner of calculating records and further it will be stored on log structure. Further, the time of the session identification is based on the clicks of every page from the website. This unique identification is known as the (*SI*) to avoid the redundant clicks of the same page was clicked by the same session user. Then the comparison is done with the previous one if the value is null in previous record because some previous click may be empty for that page click so at that time it will take the before one from the previous session to process. So the entire surfing of log structure is never required here. Then finally combinations of the frequency results are performed based on the predefined support value and results will be stored. Methodology works are only presented in this paper; the experimental results of this work will be processed in

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

further and also future work includes the comparison of various methods.

REFERENCES

[1] DeMin Dong, "Exploration on Web Usage Mining and Its Application," Intelligent Systems and Applications, ISA 2009, International workshop on IEEE, PP. 1-4, May 2009.
[2] Dongshan Xing, Junyi Shen, "Efficient data mining for web navigation patterns," Information and Soft Technology, Elsevier, volume 46, Issue 1, pp. 55-63, August 2003.
[3] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan, "Frequent pattern mining: current status and future directions," Data Mining and Knowledge Engineering, Springer, Volume 15, Issue 1, pp.55-86, August 2007.
[4] Qingtian Han, Xiaoyan Gao, "Research of Distributed Algorithm Based on Usage Mining," Knowledge Discovery and Data mining, WKDD 2009, Second International Workshop on IEEE, Computer society, pp. 211-214, January 2009.
[5] Ran Tang and Ying Zou, "An Approach for Mining Web Service Composition Patterns from Execution Logs," Web Systems Evolution (WSE) 12th International Symposium on IEEE, PP. 53-62, September 2010.
[6] Robert Cooley, Bamshad Mobasher, J aideep Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information System, Springer, Volume 1, Issue 1, pp. 5-32, February 1999.
[7] Theint Theint Aye, "Web Log Cleaning for Mining of Web Usage Patterns," Computer Research and Development (ICCRD), 3rd International Conference on IEEE, PP. 490-494, March 2011.
[8] Tomás Arce, Pablo E. Román, Juan Velásquez, Víctor Parada, "Identifying web sessions with simulated annealing," Expert Systems with Applications, Elsevier, Volume 41, Issue 4, pp. 1593-1600, March 2014.
[9] Yu-Hui Tao, Tzung-Pei Hong, Yu-Ming Su, "Web usage mining with intentional browsing data," Expert Systems with Applications, Elsevier, Volume 34, Issue 3, pp. 1893-1904, April 2008.
[10] Zdravko Markov, Daniel T. Larose, "Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage," Springer, 2008.
[11] Zidrina Pabarskaite, Aistis Raudys "A process of knowledge discovery from web log data: Systematization and critical review," Journal of Intelligent Information Systems, Springer, Volume 28, Issue 1, pp. 79-104, February 2007.