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. 5, Special Issue 1, January 2018

AN IMPROVED TECHNIQUE FOR SESSION IDENTIFICATION WITH CONSTRAINED CLUSTERING

[1]J.Umarani, [2]Dr.S.Manikandan,
[1]Research Scholar, Research and Development Centre, Bharathiyar University, Coimbatore,
Email:umashenthaan@gmail.com
[2]Head, Department of Computer Science and Engineering, Sriram Engineering College, Chennai
Email: manidindigu@gmail.com

Abstract:-

The current scenario of the Information Technology and the World Wide Web continues to grow in both the size and complexity of websites, the privacy preserving of the web access are the crucial part in the field of IT and WWW. Due to the increasing complexity of the WWW, web site publisher wish to know the users of the site and facing the problem to attracting and retaining their user's based on their needs, because of this the publisher update their websites features. Web usage mining is one of the applications in the data mining techniques to web usage log repositories in order to realize the usage patterns that can be used to analyze the user's behaviour in detail [1]. WUM consist the three main steps namely preprocessing, knowledge extraction and pattern analysis. The target of the preprocessing step in WUM is to convert the web log data into a set of user profiles. Each such profile captures a sequence or a set of IP/URLs representing a user session. From the web log file the session to be identified by the IP/Agent, Referrer and time-outheuristics. In this research the different kinds of like browser log, server log and proxy log file are taken to the analysis. First step the web log files are cleaned by the Data Cleaning algorithm it is converted to a Data Set. Constraint clustering algorithm used for analysis of sessions.

Index Terms:
Web Usage Mining, Web Log Files, Session Identification, Data Mining Algorithms, Constraint clustering algorithm.

I.INTRODUCTION

The Internet is a vast technology to hold information and sharing of information in around the world, the knowledge are information's are shared by the different kinds peoples or users. The technology consists of the Internet is ISP, Web Server, Proxy server, client, web site, web page, web user etc.,. Web Mining is one of the data mining technology to analysis the web data's, it is further divided into three categorized into three based on the data to be mined. They are Content Mining, Structure Mining and Web Usage Mining. Content Mining is the process of Extracting documents from the web. Structure Mining is the process of discovering structural information from the web. Web usage mining is the third categories of the web mining; it discovers the access pattern analysis of the web.

Web Usage Mining is adetection of attractive user access patterns from Web server logs, has become the subject of intensive research, because of its potential for personalized services, adaptive Web sites, and organization and presentation of Web sites. With the transformation of the Web into the education sectors. In the educational institutions are implemented on the all communications of the administrative side to Academic and officials are handled their works through the websites. The student teacher, student administration also monitored through the web. All transactions of entire services done by the educational institutions are monitored to the servers. The web server, client browser and proxy servers are kept transaction details in a separate file called as log files, each of them have their own formats.

Also the user restriction can be performed in The Web Server Log is an important source for performing Web Usage Mining because it clearly records the browsing activities of site visitors. It provides the communication activities of the client and server that is request made by the client and what responses provided by the server are kept in the log file. However, because of considerations of privacy, the logs, by default, do not record user ids.For meaningful Web Usage Mining, on the other hand, these requests must be identified into user sessions as semantic units of analysis. The difficulty of identifying users and user sessions from Web server logs has been addressed in by several researchers [1, 2, 3]. A solution to this problem is to create heuristics that capture in a logical way the behaviour of users and map it onto the Web logs.

Session Identification Heuristics

In web usage mining consists the different stages among the list the preprocessing is the first stage, preprocessing further it consists four steps they are data cleaning, user identification, session identification and path completion. Web log data's are preprocessed in multiple aspects, session identification is one among them.

In this research work represents the study and evaluation of sessions and session construction. Session can be defined as the set of pages visited by the same user within the duration of one particular visit to a website. In this regarding the researcher uses the composite heuristics based on the heuristics listed below;

- ✓ Hus1: To count the user sessions within a period. It is specified by the administrator.
- ✓ Hus2: To find out the sessions with the respective user based on the clickstream of each user. Each click stream of user represents the portions.
- ✓ Hus3: Session evaluation by total session time of user, to calculate the page viewing time with constant page view time set by the web page administrator.
- ✓ Hus4: Session evaluation by page stay time. Find out the difference between the two timestamps, if it exceeds 10 minutes then the second entry is assured as a new session.
- ✓ Hus5: To calculate browsing time, with the help of Browsing Minimum and Browsing Maximum.
- ✓ Hus6: Page Navigation based methods; it uses web topology in graph format. It considers web page connectivity; however it is not necessary to have a hyperlink between two consecutive page requests.If a web page is not connected with previously visited page in a session, then it is considered as a different session.

II.RELATED WORKS

The mentioned above heuristics is the essential of session identification process.In the proposed methodology section elaborately discuss with the specific heuristic approaches for this work.

The rest of this paper is organized as follows.Section 2 describestheDifferent author works are related to this research work. Section 3 is a methodology proposed by the researcher with different sub headings. Section 4 describes theheuristics combined for the evaluation. Section 5explains the methodology used for the evaluation ofheuristics. In section 6, we compare the performanceof the heuristics for the site under consideration. Section 7 proposes a strategy for further work inevaluating heuristics and concludes the paper.

In [1] the authors discussed their own methodology to preprocess the web log data including data cleaning, user identification and session identification, they are discussed the details about how to apply the Fuzzy C-Means Clustering algorithm in order to cluster the user sessions. For the improvement of the clustering results, they proposed a Fuzzy Set Theoretic approach for the removing the sessions below a specified threshold and also assign the weights to all the using a Fuzzy Membership Function based on the number of URLs accessed by the sessions. The proposed methodology performs the feature subset selection of session vectors and session weight assignment. At the final the authors compared their soft computing based approach of session weight assignment with the traditional hard computing based approach of small session elimination. The results shows that the Fuzzy Set Theoretic approach of session weight assignment results in better minimization of clustering performance index than without session weight assignment.

In [2] the authors discussed their own approaches for the Session Identifications in Click Stream Analysis, and also implement their algorithm in Java Programming Language with the NetBeans Editor. The author used the database with logs from NASA Website; it is downloaded from the online. The parameter number of session for the proposed algorithm is greater than the normal/classic algorithm, and also to avoid a division in too many sessions, the author increases the value of 300 seconds used in the modified algorithm case. The final outcomes are represented in the figures and charts of the results and discussion.

In [3] author implemented his own methodology and finally has reported on the investigation into the efficiency of heuristics that may be used to identify user sessions, based on the analysis of Web logs. This was done on the basis of knowing the 'local' circumstances, policies and procedures, which allowed for much better estimates of user sessions. The heuristics, by their very nature, are generic, without this local knowledge. It is still too early to draw definite quantitative conclusions about the efficiency of these heuristics in combination or individually. However, finally the author provides the four investigation steps to implement their thoughts they listed below.

- ✓ Start with the log data bearing in mind that it is just the raw data and needs to be 'cleaned up'.
- ✓ Formulate and use a 'Cleaning' procedure. The cleaning procedure will depend upon the local circumstances and policies. Thus, for example, the logs we analysed contained error messages, spurious information, non- HTML requests and wrong (user) ids. It is possible to use more than one log file to log these entries separately. Also, the logs may be hourly, daily, weekly or any time period as determined by the Web administrators. The cleaning procedures must incorporate such knowledge.
- ✓ If any part of the logs contains explicit user ids or session ids, use them to isolate the subsets of data where such local practices will make it easier to understand

how much the heuristics vary in their analysis.

- ✓ Analyse the remaining, cleaned data that does not contain any user/session ids, on the basis of the heuristics. Use the understanding from step 3 to refine the analysis, as necessary.

In [4] the author to fix their main goal of this research was to find some golden nuggets from the raw access logs and this has been achieved. The idea of obtaining as much information as possible about the end user from the logs is a promising and less explored form of web log analysis. There are many ways of gaining personal information from the user when a person visits a particular website with the help of cookies, registration forms, java scripts on the client and server ends. But the idea of gaining information from the user IP and then developing some feature sets from the data is novel. Also provided a detailed review of various techniques to preprocess the web log data including data fusion, data cleaning, user identification and session identification. The results from this work can be used to structure the website information according to the personal preferences of the respective user category in E-Commerce sites.

In [5] the author presented distinct user identification technique which enhancement of preprocessing steps of web log usage data in data mining. They are use two pre-processing technique combine within one pre-processing step time of user identification and find out distinct user based on their attended session time. They introduced one proposed algorithm for advanced pre-processing DUI algorithm is very efficient as compare to other identification techniques. They stated to get more precious accurate result. Based on this we can easily personalized websites, improve the design of WebPages. As usages of users on websites. They are suggests about the future work, it needs to be done to combine whole process of WUM. A complete methodology covering such as pattern discovery and pattern analysis will be more userful in identification method.

In [6] the author presented their work in the following aspects. First of all, the will take the log file for processing of their view, it often contain enormous data which require a significant amount of time to be processed. Session identification is done on the time spent each page of web pages assessed by an individual user in a particular session. Two clustering techniques are performed on the preprocessed log data. Clustering is done to find frequently accessed web pages the file. When examining k-means clustering algorithm with farthest first clustering algorithm shows significant improvement in execution time performance. By implementing these techniques, faster analysis of the log file can be done even though the amount of data might be high. Reorganizing website based on the frequently accessed webpages using data algorithm is considered as a future direction of research. Reorganizing website is mainly to reduce page access delay and to provide desired information in a fewer click.

In [7] the author presented a research work for the session identification as the following sequences of stages, the main focus of the author is to compares four basic session identification approaches such as two time oriented i.e. session duration and page stay and two referrer oriented heuristics i.e. using referrer information and extended referrer heuristics by performing experiments on data collected from Banaras Hindu University Website server. From experimental analysis it has been observed that Page stay heuristic works better than other three approaches as it generates small number of short sessions and optimal number of medium and long sessions. Apart from that, we also have observed the impact of robot session on session count and session length and analysed that discarding robots session affects more for referrer oriented heuristics because robots session may contain more number of null referrer than user session. Referrer oriented heuristics could do better if a potential robot detection approach would be applied which could result in fewer number of short sessions.

In [8] the author presents a own methodology for the pre-processing in Web Usage Mining his views are represented as follows, Web log mining is one of the recent areas of research in Data mining. To use the web usage mining efficiently, it is important to use the pre-processing steps. Steps of pre-processing are analyzed and tested successfully with sample web server log files. This work delivers the steps of pre-processing including data cleaning, user identification, session identification and path completion. Once pre-processing is performed on web server log, then patterns are discovered using data mining techniques such as statistical analysis, association, clustering and pattern matching on pre-processed data, then the discovered patterns are analyzed for various applications such as web personalization, site improvement, site modification, business intelligence, etc.

Table:1. Summary of the related works used in this work

S.No	Author Name	Functionality focused by the author	Merits	Demerits
01	Zahid Ansari et.al	Proposed a new method Fuzzy Set Theoretic approach for the removing the sessions	Better minimization of clustering performance index	Session weight assignment not properly used
02	Dumitru Ciobanu et.al	Session Identifications with Click Stream Analysis	The parameter number of session for the proposed algorithm is greater than the normal/classic algorithm	More time taken to execute the algorithm
03	Amithalal Caldera et.al	The heuristics are build to attain the goal and implemented	Newly constructed heuristics are consider the existing concepts	Execution time increase but got more accuracy
04	Dr.P.Sengottuvelan et.al	Extract the essential information from the log files with the use of Data Cleaning Algorithm	Preprocessing steps are done an easy way	Session Identification done by user categories
05	Ashwin G et.al	Distinct user identification technique	To produce precious accurate result	To suggests the all preprocessing stages are altogether
06	S.Kalivani et.al	Uses the K-means and Farthest algorithms	Reorganizing website is mainly to reduce page access delay	Algorithms take more time
07	Mitali Srinivastava et.al	Referrer oriented heuristics	Potential robot detection approach is used	Robots session may contain more number of null referrer than user session
08	S.Prince Mary et.al	Clustering and pattern matching	Web personalization with the data mining	Elaborately discussed only on user identification
09	Pablo E.Roman et.al	Sessionization using bipartite cardinality matching (BCM) and integer programming (SIP)	It explore the likelihood of specific sessions and characteristics of sessions	Size and number of sessions are monitored
10	Don Koch et.al	URL rewriting can also be used to track users	It provide more accurate data	re-engineering effort

In [9] the author presented a new approach for sessionization using bipartite cardinality matching (BCM) and integer programming (SIP). And also test their approach using real sessions retrieved from an academic web site over 15 months. They are compare real sessions, results obtained by their optimization models, and results from a commonly-used timeout heuristic. They are finding their optimization models dominate the timeout heuristic using several comparison measures. For example, SIP has precision of 77:9%, BCM has nearly the same precision 77:8%, and the timeout heuristic is a distant third with only 50:9%. They are also provided variations of our optimization models to further explore the likelihood of specific sessions and characteristics of sessions. Specifically, find the maximum number of copies of a given session, the maximum number of sessions of a given size, and maximum number of sessions with a given web page requested in a given position for each session.

In [10] the authors provides their views in the aspects of tracking users and building sessions as follows web server logs present a unique challenge for web mining. They are however a ubiquitous, cheap and somewhat standard source of data to mine your e-commerce channels.With a good understanding of how your site works it is possible to properly prepare your web logs for data mining. At a minimum server generated user tracking cookies should be employed. If a web application environment is employed then URL rewriting can also be used to track users. These changes are very unobtrusive to a site and can reap large benefits in analysis. Further reengineering of web site and web applications can provide more accurate data. The principles laid out in this paper can be used in any re-engineering effort aimed at better understanding of your on-line customers.

Table 1: Demonstrates the summarizes of the respective related work presented by the different author provides their own view regarding the session construction in preprocessing steps of web usage mining all of them have some deficiency, to overcome that the researcher proposed a new approach for session construction. The are elaborately described in

III. DATA COLLECTION

In web environment transactions are kept in the web log files, its mainly classified in the three categories namely 1).Web Server Log 2).Proxy Log 3).Client/Browser Log [11].

1.Web Server Log File

The most significant and frequently used source for WUM is web server log data. The web log data is generated automatically by web server when it services user request and response, which contains all information about visitor's activity request [11]. The common server log file types are Access log, Agent log, Error Log and Referrer log [11] the following table represents the server log file type.

Table 2: Web log file types

Log File Type	Content of Log file
Access Log	All resource access request sent by user
Agent Log	User's browser, version, Operating System etc
Error Log	Details of errors occurred while processing user access request.
Referrer Log	Contains information about referrer page.

Web log files data are vary depends on the web server format and number of attribute available in the log file. The categories of log files are as follows; 1).Common Log Format 2).Extended Common Log Format 3).Centralized Log Format 4).NCSA Log file format 5).ODBC Logging 6).Centralized Binary Logging.

Following is an example line of access log in common log format.

```
121.546.87.7-[25/Apr/2017:01:03:31          -0500]
"GET/HTTP/1.0" 200 3270
```

This line consist the following fields.

Client IP address , User id ('-'if anonymous), Access time, HTTP request method, Path of the resource on the Web server, Protocol used for the transmission, Status code returned by the server,Number of bytes transmitted. [12] discussed about a method, Wireless sensor networks utilize large numbers of wireless sensor nodes to collect information from their sensing terrain. Wireless sensor nodes are battery-powered devices. Energy saving is always crucial to the lifetime of a wireless sensor network. Recently, many algorithms are proposed to tackle the energy saving problem in wireless sensor networks. There are strong needs to develop wireless sensor networks algorithms with optimization priorities biased to aspects besides energy saving. In this project, a delay-aware data collection network structure for wireless sensor networks is proposed based on Multi hop Cluster Network. The objective of the proposed network structure is to determine delays in the data collection processes. The path with minimized delay through which the data can be transmitted from source to destination is also determined. AODV protocol is used to route the data packets from the source to destination.

2. Proxy Log

Enterprises or companies can have own server for handling internet services by the use of a dedicated machine known as a proxy server, all the request and response are serviced through this proxy server. Study of this proxy server log files, whose format is same as of web log file may reveal the actual HTTP requests coming from multiple clients to multiple web servers and characterizes, reveals the browsing behavior for a group of anonymous users sharing a common proxy server [11]. Some web sites use n-tier architecture to have reliable, efficient and secure web applications. Log data that are gathered at application server while servicing the users request can also be used for web usage mining. They peculiarly show\how user requests are serviced and may assist in identifying

3. Client/Browser Log

The client machine uses any one of the browser like Firefox, Google chrome or Opera etc., the browser recording the activities, events that happens within the location of client machines. Like mouse wheel rotation,scrolling within a particular page, mouse clicks, contentselectionThere are number of ways is used to associate the application to the internet, it also maintain their client/browser log details.

3.1 SOURCE DETAILS

The Web logs are collected from the Internet Server which is located at Thanthai Roever Group of Educational Institutions, In that institutions have a one centralized server and four proxy servers, these are located in the respective institutions among the group. And also few of client logs are also collected from the client machines connected to the proxy servers. The following are the server's details;

1. RMS01-RoeverMainServer(198.162.1.100)
2. RMS02-RoeverMainServer(198.162.1.101)
3. RECPROXY-(198.162.2.100)
4. THRCPROXY-(198.162.3.100)

Table 3: Log files Server description with No.of. Records

S.No	Server Description	No.of. Records
01	RMS01-RoeverMainServer(198.162.1.100)	17000
02	RMS02-RoeverMainServer(198.162.1.101)	26000
03	RECPROXY-(198.162.2.100)	57500
04	THRCPROXY-(198.162.3.100)	62000
05	Client log file from client machines	41000
	Total Number of records	203500

The academic year of college is started from June month of the every year. The above mentioned records are collected from these servers on the period of September 2017and October 2017. Table: 3Provides the details of log servers and number of records collected from the respective servers.

IV. PROPOSEDMETHODOLOGY

CONSTRAINED SESSION CONSTRUCTION CLUSTERING ALGORITHM

4.1 Architecture of the CSCCA

The preprocessing steps of web usage mining consists the data cleaning, user identification, session identification/construction and path completion. The previous works of this research represents the data cleaning and user identification. This work represents the session identification/construction, the following figure consists the different stages. They are 1).Data Cleaning 2). Target Data Storage 3). Data Set Preparation 4). Constrained Session Construction Clustering Algorithm, the fourth stages consists the sub sections are as follows; a). Session Construction with IP b). Session Construction with Timeout Heuristics c). Session Construction with Referrer Heuristics. 5). Session data's withECBCAA 6). Comparative Analysis of the ECBCAA and CSCCA.

STAGE 1: DATA CLEANING

The collection of web log informations are collectively stored into the single file, such file having a nosie data based on the user requirements the logfiles are cleaned. The image files not taken for the cleaning process, if the reading process find out the image file extensions, they are neglected by the reading process and also if find the error code and system calls they also neglaeted.

STAGE 2: TARGET DATA STORAGE

The final steps of the data cleaning process are having cleaned web log information it consists the different attributes like the IP Address, date, time, and Method GET and POST. This information's are stored into the target databases; it contains the high volume data.

STAGE 3: DATA SET PREPARATION

The proposed algorithm have the more parameters, its are extracted from the target data storage and generate the data set based on the rule applied to the proposed algorithm. The information's or data are kept in the data set are used for the different algorithm used for this proposed methodology.

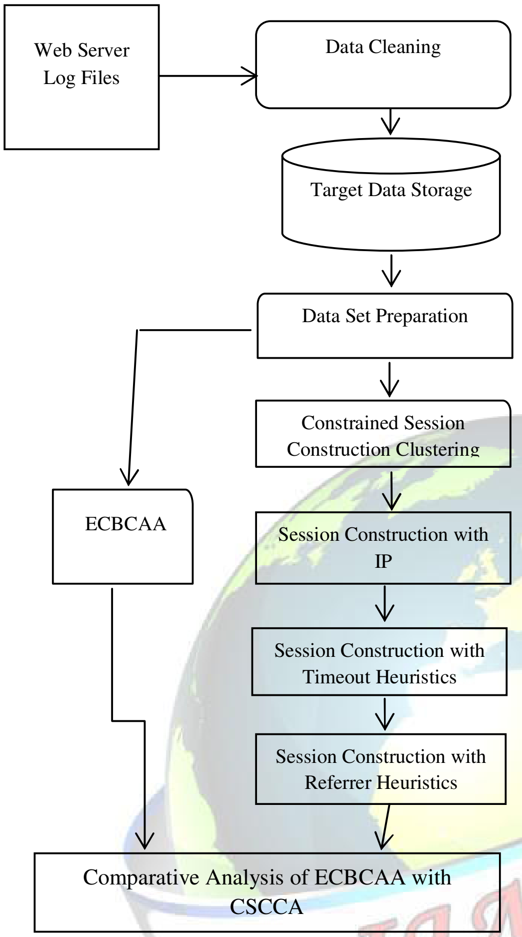
Web Server Log Files
Data Cleaning
Target Data Storage
Data Set Preparation
Constrained Session Construction Clustering
ECBCAA
Session Construction with IP
Session Construction with Timeout Heuristics
Session Construction with Referrer Heuristics
Comparative Analysis of ECBCAA with CSCCA

Figure:1.**Architecture of the CSCCA**

Figure 1: Represents the complete architecture of the CSCCA and workflow of the different stages of the constrained session construction clustering algorithm.

STAGE 4: CONSTRAINED SESSION CONSTRUCTION CLUSTERING ALGORITHM

The proposed methodology implemented through the following core algorithm with clustering concepts, the constrained condition can apply with the following options which consists the three case options that chooses the any one of them, then it call the appropriate algorithm and execute them return to output. The following Algorithm: 1. represents the main algorithm which constrained session construction clustering Algorithm.

Algorithm: 1.Constrained Session Construction Clustering Algorithm

```
Input: Pre-processed web log data in Relation Rn, object ol.
Output: Session clusters
```

```
Step 1: string obinput;
Step 2: assign ol1=IPAH, ol2=TMOH, ol3=RFRH;
Step 3: Switch (obinput)
Step 4: begin
Step 5: case 1:
Step 6: Call (IPAH);
Step 7: Break;
Step 8: case 2:
Step 9: Call (TMH);
Step 10: Break;
Step 11: case 3:
Step 12: Call (RFRH);
Step 13: Break;
Step 14: Default;
Step 15: Bad input given by the user;
Step 16: Break;
Step 17: end.
```

STAGE 4.1: SESSION CONSTRUCTION WITH IP

In this algorithm it takes the input from the relation Rn and with the object ol, it consists the attribute which is the IP Address, when the ipaddck_id Conditions checked with the range between (198.162.2.100 and 198.162.2.255) and (between 198.3.100 and 198.162.3.175) then call the timeout heuristics and execute them return the output. Otherwise call the RFRH algorithm and execute them and return the value. The Algorithm: 2. Represents the IP Address Validation and execute them with the sequence of steps.

Algorithm:2. IP Address validation (IPAH)

```
Input:Preprocessed web log Relation Rn,ol (ol1-IPA)
Output:session result
Step1: read ipadd_id;
Step2: assign ipaddck_id=0;
Step3: if ipaddck_id =(between 198.162.2.100 and 198.162.2.225) and (ipaddck_id = (between 198.162.3.100 and 198.162.3.175)) then
Step4: call (TMH); return ();
Step5: else
Step6: call (RFRH); return ();
Step7: endif.
```

STAGE 4.2: SESSION CONSTRUCTION WITH TIMEOUT HEURISTICS

The second sub stage of proposed methodology construct the session with time based heuristics, here the threshold value of times fixed as follows; the page minimum time as 3 minutes, the page maximum time is 30 minutes and page stay time as 15 minutes based on this values the session usage time calculated from the web log files. To calculate the page waiting time as subtract the page starting time from the page end time, if the page waiting time met the page minimum and page maximum threshold value then count the session subsequently do the process up to the page waiting time is to reach the value zero or page minimum value.

Algorithm: Timeout Heuristics (TMH)

```
Input:Preprocessed web log Relation Rn, object ol2 (TMH)
Output: Total Session Time (TST-Session list), Page Stay Time (PST-Session list )
Step1: begin
Step2: assign pmint=3 minutes,pmaxt=30 minutes, pst=15 minutes, sess_count=0;
Step3: for i=$r_1$ to $r_n$ step 1
Step4: begin
Step5: read user_id, pstart_time,pend_time from the ol2;
Step6:pvt=pend_time-pstart_time;
Step7: if (pvt>pmint) and (pvt<=pmaxt ) then sess_count=sess_count+1
Step8: else pvt=pvt-30;
Step9: repeat step6and step7 until pvt=0;
Step10: end
Step11: if (user open a page but no further movement) then (count the page waiting time pwt)
Step12: if pwt=pst then terminate user session; sess_count=0;
Step13: endif
Step14: endif
Step15: end
```

STAGE 4.3: SESSION CONSTRUCTION WITH REFERRER HEURISTICS

The third sub stage of proposed methodology construct the session with referrer based heuristics, here the threshold value of time T fixed as 10 minutes, the reference page R1 session is **S** and the reference page R2 session is S **S,** then calculate the session of R1 and R2. In the other option of session counting is referrer with timestamp in this work fixed the timestamp value as 10 minutes, then count the session of R1 and R2 are time stamp of R1 and R2 exceed the 10 minutes then count the session as one. Subsequently do the process up to referrer value is null. Return back result to the main algorithm.

Algorithm: Referrer Oriented Heuristics

```
Input Preprocessed web log Relation Rn, object ol3 (RFRH)
Output:Referrer Session Result
Step1: assign R1=null, R2=null, T1=0, T2=0, tcons=10 minutes;
Step2: if (R1 refer R2 without timestamp) then sessionR1=S; sessionR2=SS;
return sessionR1,return sessionR2;
Step3: else if (R1 refer R2 with timestamp T1 and T2) then sessionR1=ST; sessionR2=SST;
T=T2-T1;
Step4: else if T<=tcons then return T;
Step5: else return T2;
Step6: endif; endif; endif;
Step7: end.
```

V.EXPERIMENTAL STUDY

Table 4:.Log files Status

S.No	Description	No.of. Records
01	Total Number of record in original log file	203500
02	After Removing the entries which are not having status code	201000
03	Total No.of Attributes in original log file	13
04	Total No.of Attributes in Cleaned log file	8

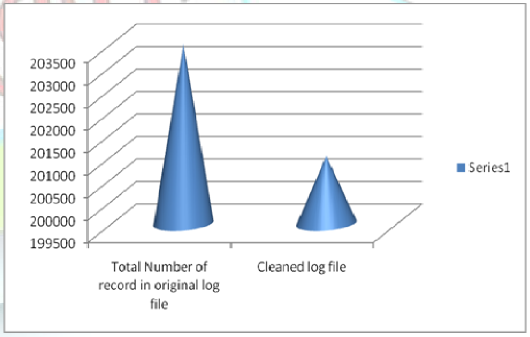

Figure: 2. Log file status graphical representation

The table 4 describes the input that is number of records from the server log file and it is taken to the data cleaning process after that displays the number of records for further process. The figure 2 also represents the graphical representation of the log files.

Table 5:. Session Type and No.of. Session in log file

S.No	Session types	No.of. Sessions
01	IP Based Sessions	90300
02	Time Based Sessions	90600
03	Referrer Based Sessions	20100
	Total Sessions	201000

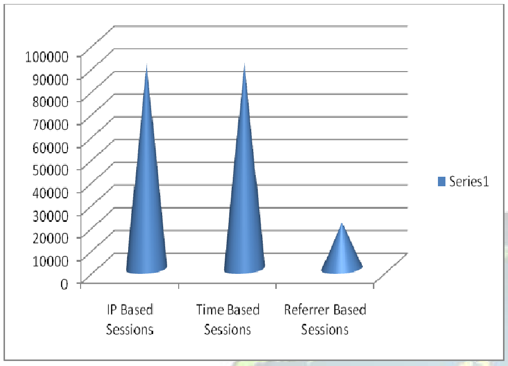

Figure 3: Session Type and No.of. Session in log file

Table 4: Time Comparison of the two algorithms implemented for session constructions

S.No	Algorithm	Size of file	Time complexity
01	**ECBCAA**	120 KB	0.3 seconds
02	**CSCCA**	120 KB	0.22 seconds

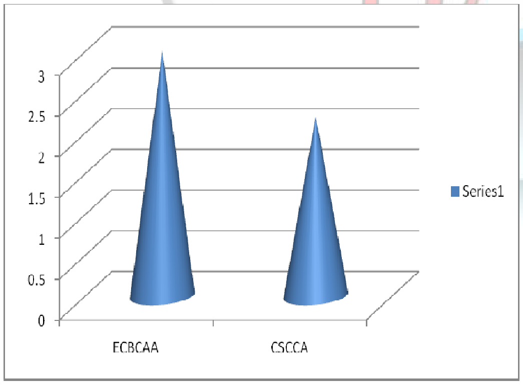

Figure 6: Time Comparison of the two algorithms implemented for session constructions

VI. CONCLUSION

In web usage mining consists the different stage the preprocessing is one among them, in that it have different stages the session construction is one among them, session identified in web usage mining in different aspects by the authors view here the researcher uses the constrained clustering algorithm with the integrated heuristics approach to found are identify the construction. It also compared with the another algorithm with the time complexity this algorithm take the minimum time and also with the high efficiency

REFERENCES

[1] Zahid Ansari, Mohammad Fazle Azeem, A. Vinaya Babu and Waseem Ahmed, A Fuzzy Clustering Based Approach for Mining Usage Profiles from Web Log Data, (IJCSIS) International Journal of Computer Science and Information Security,Vol. 9, No. 6, 2011, PP:70-79.

[2] Dumitru Ciobanu, Claudia Elena Dinuca, A New Method for Session Identification in Clickstream Analysis, Recent Researches in Tourism and Economic Development,Recent Researches in Tourism and Economic Development, ISBN: 978-1-61804-043-5, pp-476-479.

[3] Amithalal Caldera and Yogesh Deshpande, Evaluation of Session Identification Heuristics in Web Usage Mining, School of Computing and Information Technology, College of Science, Technology and Engineering University of Western Sydney PO Box 1797, Penrith South DC, NSW 1797, Australia.

[4] Dr.P.Sengottuvelan, R.Lokeshkumar, D.C.Karthikkumar and R.Sinduhuja, Session Identification in Web Usage Mining to Personalize the Web, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 9 (2015), © Research India Publications http://www.ripublication.com.

[5] Ashwin G. Raiyani, Prof. Sheetal S.Pandya, Discovering user Identification Mining technique for preprocessed Web Log Data, Journal of Information knowledge and Research in Computer Engineering.

[6] S.Kalivani and K.Shyamala, Clustering of Web users behavior based on the Session Identification through the web server log files, International Journal of Control theory and Applications, ISSN: 0974-5572.

[7] Mitali Srinivastava, Atul Kumar Srivastava, Rakhi Garg and P.K.Mishra, Experimental Study of Time oriented and Referrer oriented Session Identification Methods in Web Usage Mining, IJEE, Volume 9, Number 01 Jan -June 2017 pp. 177-183.

[8] S.Prince Mary and E. Baburaj, An Efficient Approach to Perform Pre-Processing, Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166 Vol. 4 No.5 Oct-Nov 2013,PP:404.

[9] Pablo E.Roman, Robert F.Dell , D.Velasquez and Pablo S.Loyola, Identifying User Sessions from Web Logs with Integer Programming, California 93943, USA.

[10] Don Koch, John Brocklebank and Tom Grant Richard, Mining Web Server Logs: Tracking users and Building Sessions, paper 154-27,SUGI 27, Data Warehousing and Enterprise Solutions.

[11] Suneetha, K. R. and D. R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", (IJCSNS) International Journal of Computer Science and Network Security, VOL.9, No.4, April 2009.

[12] Christo Ananth, T.Rashmi Anns, R.K.Shunmuga Priya, K.Mala, "Delay-Aware Data Collection Network Structure For WSN", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Volume 1,Special Issue 2 - November 2015, pp.17-21