The Research of Preprocessing and Pattern Discovery Techniques on Web Log files

P.Dhanalakshmi
Research Scholar
Computer Science & Engineering
JNTUA,Anantapuramu
mallidhana5@gmail.com

Dr. K.Ramani
Professor & Head
Department of IT
SVEC, A.Rangampet
ramanidileep@yahoo.com

Dr. B.Eswara Reddy
Vice Principal
JNTUA, Anantapuramu
eswarcsejntu@gmail.com

***Abstract*—The increased on-line applications are leading to exponential growth of the web content. Most of the business organizations are interested to know the web user behavior to enhance their business. In this context, users navigation in static and dynamic web applications plays an important role in understanding user's interests. The static mining techniques may not be suitable as it is for dynamic web log files and decision making. Traditional web log preprocessing approaches and weblog usage patterns have limitations to analyze the content relationship with the browsing history This paper, focuses on various static web log preprocessing and mining techniques and their applicable limitations for dynamic web mining.**

keywords—Static Logs,Graph models, Asscociation rules, Web log, Navigation patterns.

I. INTRODUCTION

The personalized web recommendation system is becoming increasingly important due to its high utility and the availability of a large number of web page content. Many researchers have tried to implement online recommended systems by using static behavior models. Static models are used to describe the user's short term profile by using user's web request. The nature of static problems is associated with the historical data, i.e. the lack of interaction between a product and the user and between the two or more products. The existence of missing data in dynamic web content seems to be more significant than the static data. However, to analyze the dynamic web content, the recommended system needs to parse a lot of historical data and predict, how the customer will browse the page or web product [1].Web log mining is the process of analyzing user behavior and user navigation patterns in static web logs or dynamic web logs. The majority of the web customers are non-experts and find it difficult to study the historical user's patterns and their behavior towards the online content. Moreover, the emergence of online services such as e-commerce, e-banking and e-learning has changed the purpose in which turning web sites into businesses and increasing the business competition[2].

Sequential pattern mining models are applied to discover the frequent web usage patterns between the page requests, session time and browsing history, etc. However, these sequential models have certain limitations such as:

- Need to maintain huge data structure in memory space throughout the execution due to the database scans.
- Increase in memory size due to its high dimensional attributes and values.
- Lack of predicting a user's next access patterns based on historical data.

Web usage mining applications are used to find the web visitors' profiles and their behavior in terms of strengths and weaknesses of their web applications. The main issue focused by any web usage model is data increases per second with different server log file formats. Learning about the customer's behavior , predict their requirements in the future , monitoring the file structure and content of the web service according to their navigation behavior is necessary. Accurate web usage patterns could help to improve the new users, retain existing customers, optimize cross sales, customers' interest, etc. The usage decision patterns can improve the web server efficiency by using different caching techniques so as to minimize the server response time. The user's profile could be designed by integrating customer's page navigation paths with other attributes such as server response, session time, page duration, hyperlink and page content.

Applications of web usage mining include mining conceptual visiting user profile hierarchies and interesting patterns from the web log files for building the frequent web access structures using tree based Markov model or association models. Since web usage mining approaches consider only server logs due to security issue of information on the client side. The set of limitations of the server side are :

- IP addresses and sequence of page requests in the log file are not a reliable fields, because some pages are cached by the web server or browser and proxy.
- It is difficult to interpret the session duration in the server log file, as the same IP address can be used different users at different intervals (i.e. 30 minutes default time).
- Also server log files are difficult to predict without log preprocessing.
- Since server log files have different structures and formats, it is difficult to apply same preprocessing or knowledge based techniques.

A. Static Web Pattern Mining:

The basic structure of the static web log framework has four phases namely static data collection, data cleaning , pattern discovery techniques and pattern analysis with output. This framework can be shown in Fig.1. In the first phase, static web log files are extracted from the server in one of the standard formats using temporal basis. Since, the server log files are raw data with uncertain information, it is preprocessed using field extraction, user identification and session

978-1-4673-8286-1/16 $31.00 © 2016 IEEE
DOI 10.1109/IACC.2016.35

identification algorithms. Since the server log format and the number of fields are fixed, this process is repeated to all the server logs in the given time interval. If the log file format is changed, then the whole preprocessing steps are to be modified according to the input log format.

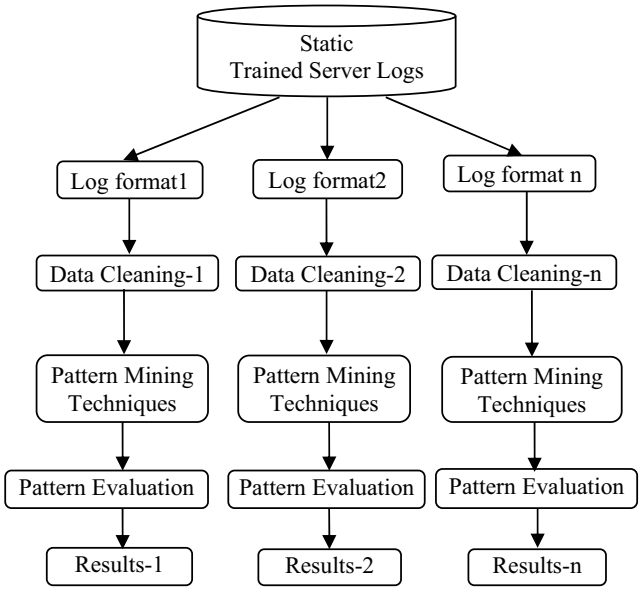

Figure 1. Static Log File Usage Evaluation

In the next phase, pattern mining techniques such as association mining, clustering, prediction and classification models are evaluated to discover the interesting patterns on the preprocessed log files. In the pattern evaluation phase, each pattern is tested on the newly preprocessed log files for validation and then results are displayed to the user.

The rest of the paper is organized as follows. Section II describes the related work of the static web mining techniques such as preprocessing and pattern discovery techniques and its limitations. In Section III , we have discussed different static models on server log files and its performance is evaluated on the different periodic log files. In Section IV, experimental results of different static models are evaluated and finally, SectionV describes about conclusion and future scope.

II Related Work

In [3], Kun Chang Lee and Soonjae Kwon implemented the web recommender system on server log files, but the problem with this system is that they are highly dependent on user requests, session time and available number of resources. Another issue is that if a user is likely to view sports news in the morning and business and financial news in the evening then existing recommended systems are lacking in providing temporal based recommendations. Graph based user interests are evaluated in [4] , to find the product inter and intra relationship between the products. Different graph based clusters along with product association are shown in Fig.2.

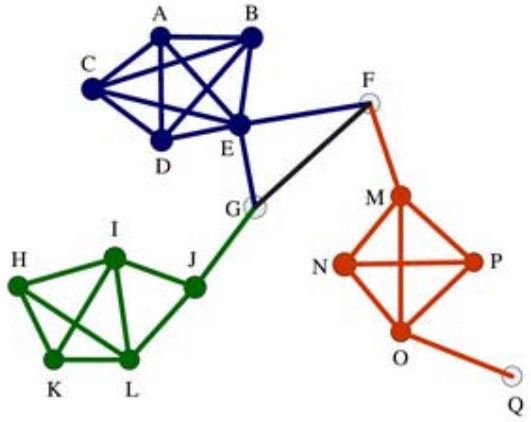

Figure 2.: Graph based Product Recommended System

Nisarg Pathak et al., [5] Proposed a model for navigation pattern discovery based on the association rules between the user accessed web pages. First, they have collected log files from the web server and then expectation maximization and DBSCAN algorithms are applied to cluster the similar features within the log file. Finally, navigation patterns are analyzed using the clustered results.Its limitations is, it considers only static log files and need to improve new user navigation patterns for dynamic websites.

K. Sudheer Reddy et al., [6] implemented the preprocessing method to filter web log files from the server. They have taken log files from different sources to filter and analyze relevant information before applying the pattern mining algorithms. In this system they applied data cleaning algorithm on 18 attributes out of which 17 attributes are found irrelevant for pattern discovery. The main limitations of this model are, need to optimize the preprocessing for huge data sets with high dimensional data and to optimize the field extraction algorithm for longer sessions.

Maheswari, B.U et al., [7] Proposed a new model session identification and page reconstruction process in server log files. They implemented user profile clustering technique on the preprocessed data to find the similar relationships between the user profiles. The main issue in this clustering model is scalability and the ability to handle large number of log files.

Vincent S. Tseng et al., [8] proposed a graph based model to find the user's web access behavior and frequent sequential web access patterns.They used two phases, one is input logs are preprocessed to get the list of accessed request pages in different timestamps. In the second phase, web usage patterns are represented using graph model. Finally, sequential access patterns are extracted for decision making. The main limitation of this model is, as the web log size increases, graph size and access patterns also increase.

Yan Li,Boqin Feng et al., [9] implemented a path completion algorithm in web usage mining. A path is the collection of web pages traversed in one or more sessions. The main objective of the path completion

model is navigation analysis of log files. This model fails to extract the inter and intra based web log patterns.

The basic path construction process has the following steps:

Path construction:
Read the url list in Urldata
If Urldata[i] !=null ;
If Urldata[i] == Urldata[1]
Continue;
Else
Get the web user request details such as link name, session id, request bytes , agent name etc.
Construct the graph nodes to be traversed between the Urldata[1] and Urldata[i].
Repeat the steps until i=2,3…n
Extract Path construction patterns from source node to destination nodes.

Baoyao Zhou et al., [10] designed a model to predict the web usage patterns using different temporal time periods.Proposed a framework for predicting time based navigation patterns which consist of calendar schema for representing time parameters and the time period based N-gram model for predicting user's temporal navigation patterns. The temporal model doesn't consider the prediction on group user navigation patterns for pattern evaluation and validation.

Xidongwang [11] proposed a static model that can find user's frequent access patterns in historical web data.They implemented the concept of web access patterns according to the user's navigation paths and then frequent access pattern mining algorithm.Frequent access pattern is based on FP-Tree algorithm to extract access patterns for decision making.

III. Static Pattern Discovery Model

In this section, we have implemented a static pattern discovery model on server log files for web usage mining. The main approach of the static model is to predict the web usage patterns using a sequence of four steps i.e static data collection, static log preprocessing, pattern discovery technique and pattern evaluation.

Static log data collection: Online applications are designed and developed in different programming environments and data structures.Also, data is stored in different formats and structures in web servers.Web usage applications are based on data collected from three sources such as : web clients, web servers and proxy servers. Web servers are the most common source of static log files stored in different formats and fields. They can collect huge amounts of data in their log files per unit time.These log files contain fields such as IP, date and time,userrequest,source bytes, agent name, destination bytes, cookies, requested page, response code, request status,url etc. Also this information is stored in different formats such as common log format, extended log format, XMLand LOGML.

A Description of the static log instance

203.109.106.200 - - [31/May/2015:03:49:20 +0000] "GET /results/jquery.js HTTP/1.1" 200 72174

203.109.106.200 :IP address of the requested page server or IP address of the user.
[31/May/2015:03:49:20 +0000] : Date of the requested page and page access time.
GET : Page access method(either GET or POST)
/Main/cron/run.php: Relative URL of the requested page.
HTTP/1.1 : Protocol and version.
200 : Server Response code
72174: Page visit's time duration

B) Sample Static log data:

Web access logs are the files that record the users browsing information on the server. The general Apache log file is used in this paper for static models. The sample static log files are used in this paper for a period of 3-months.

66.249.65.246 - - [30/May/2015:20:33:52 +0000] "GET /LEFT_MENU.css HTTP/1.1" 200 1667
66.249.65.239 - - [30/May/2015:20:33:53 +0000] "GET /xsp_styles.css HTTP/1.1" 200 7579
66.249.65.246 - - [30/May/2015:20:40:27 +0000] "GET /Products.html HTTP/1.1" 200 10617
66.249.65.246 - - [30/May/2015:20:42:30 +0000] "GET /DataMine.html HTTP/1.1" 200 10068
66.249.65.246 - - [30/May/2015:20:56:54 +0000] "GET /Nsecurity.html HTTP/1.1" 200 10069
66.249.65.246 - - [30/May/2015:20:59:12 +0000] "GET /ImageProcess.html HTTP/1.1" 200 10072
199.79.62.54 - - [30/May/2015:21:08:02 +0000] "GET /main/cron/run.php HTTP/1.0" 200 73
66.249.65.246 - - [30/May/2015:21:15:36 +0000] "GET /SEMANTICMINING.html HTTP/1.1"
180.76.15.148 - - [30/May/2015:21:36:13 +0000] "GET /ocart/index.php? route=product/product &path=25_29&product_id=51&sort=pd.name&order=DESC&limit=50 HTTP/1.1" 200 18164
199.79.62.54 - - [30/May/2015:22:08:01 +0000] "GET /main/cron/run.php HTTP/1.0" 200 73
180.76.15.20 - - [30/May/2015:22:20:30 +0000] "GET /ocart/index.php? route=information /contact HTTP/1.1" 200 10674
180.76.15.9 - - [31/May/2015:01:40:35 +0000] "GET /ocart/index.php ?route=product/ product &manufacturer_id=11&product_id=56 HTTP/1.1" 200 17096
203.109.106.200 - - [31/May/2015:03:49:21 +0000] "GET /results/css/images/side2.jpg HTTP/1.1" 200 37393
203.109.106.200 - - [31/May/2015:03:49:35 +0000] "GET /results/css/images/side1.jpg HTTP/1.1" 200 6899
203.109.106.200 - - [31/May/2015:03:49:36 +0000] "GET /results/css/images/logo.jpg HTTP/1.1" 200 3848
203.109.106.200 - - [31/May/2015:03:49:20 +0000] "GET /results/jquery.js HTTP/1.1" 200 72174
203.109.106.200 - - [31/May/2015:03:50:20 +0000] "GET /results/css/images/banner.jpg HTTP/1.1" 200 73928
203.109.106.200 - - [31/May/2015:03:50:55 +0000] "GET /favicon.ico HTTP/1.1" 200 2462
199.79.62.54 - - [31/May/2015:04:08:07 +0000] "GET /main/cron/run.php HTTP/1.0" 200 73

199.79.62.54 - - [31/May/2015:05:08:02 +0000] "GET /main/cron/run.php HTTP/1.0" 200 73

C) Sample Dynamic data stream:

The size and nature of the dynamic log files can make session identification and pattern discovery time-consuming.Also, the data generated in the dynamic web stream consists of user's click information on the particular website and webpage navigation flow.The format of the dynamic raw file is shown below

1. localhost:8084/DynamicWebData/index.jsp
2. localhost:8084/DynamicWebData/clickstreams.jsp
3. localhost:8084/DynamicWebData/viewstream.jsp?sid=7E1C2385D7971C103F0FE3C341005EF7
4. localhost:8084/DynamicWebData/viewstream.jsp?sid=DD53F95F47D6CF3EFABD63B73FB75ECF
5. localhost:8084/DynamicWebData/streams
6. localhost:8084/DynamicWebData/streams?sid=DD53F95F47D6CF3EFABD63B73FB75ECF&showbots=false
7. localhost:8084/DynamicWebData/index.jsp
8. localhost:8084/DynamicWebData/streams
9. localhost:8084/DynamicWebData/streams?sid=DD53F95F47D6CF3EFABD63B73FB75ECF&showbots=false
10. localhost:8084/DynamicWebData/viewstream.jsp?sid=DD53F95F47D6CF3EFABD63B73FB75ECF
11. localhost:8084/DynamicWebData/
12. localhost:8084/DynamicWebData/utilities/style.css
13. localhost:8084/DynamicWebData/utilities/css
14. localhost:8084/DynamicWebData/utilities/css(1)
15. localhost:8084/DynamicWebData/utilities/font-awesome.css
16. localhost:8084/DynamicWebData/utilities/anim.css
17. localhost:8084/DynamicWebData/utilities/style1.css
18. localhost:8084/DynamicWebData/utilities/css(2)
19. localhost:8084/DynamicWebData/font-awesome.css
20. localhost:8084/DynamicWebData/fontawesome/css/font-awesome.min.css

Initial Referrer:
Hostname: 0:0:0:0:0:0:0:1
SessionID: DD53F95F47D6CF3EFABD63B73FB75ECF
Bot: No
Stream Start: Sun Nov 08 08:28:21 IST 2015
Last Request: Sun Nov 08 08:29:30 IST 2015
Session Length: 1 minutes 8 seconds
# of Requests: 53

D) Static log preprocessing:

Each static weblog file may contain unnecessary fields and values, the data cleaning step should be performed on the log files. This can be performed by checking the user's request URL and response URL of the log file. The suffix of the requested URL represents the extension of the page information such as .jsp,.html,.php,.aspx,. jpg,.jpeg,.mp3,.flv,.mp4,.png,.etc. In the static log cleaning approach, we can filter the web page extensions by removing the media and document files.Also, this approach will remove the empty log files or log files with empty fields or empty values, status codes other than 200 or 304.Finally, the cleaned static log file is ready for the user and session identification phase.

Static log cleaning Method:

Read weblog $W = \{w_1, w_2, w_3 ..... w_n\}$; of n instances

$w[i] = Set_i(m) < IP, date,$ time, request url,response url, agent,bytes...>; m fields

For each instance i in W

For each field f in w_i

If(w_i (f)==”request url” or “response url”)

then

if(w_i (f).contains(“([^\s]+(\.(?i)(jpg|png|gif|bmp))$)”)

then

remove w_i (f)

else if(“([^\\s]+(\\.(?i)(txt|doc|csv|pdf))$)”)

remove w_i (f)

else if(“([^\\s]+(\\.(?i)(mp3|mp4|flv|avi))$)”)

remove w_i (f)

else

$W'[] = w_i(f)$

End if

End if

End for

End for

E) Static User and Session Detection

In this step, different users accessing the web pages and resources are identified and filtered.This can be done by checking the IP address of the requested URL in the cleaned static log file.For each different IP field, there exists a unique user within the specified session.But if the IP address already exists in the log file then a combination of IP field and OS agent field are checked to find the unique user browsing details. A user session is a set of web pages visited by the same user within the duration of one visit to the web application.Various methods have been proposed in the literature to discover the session pages and users based on time and context. As long as the user is connected to the application, it is called the session of that user. In most of the web applications, 30 minutes timeout was used as a session duration. In order to find the session duration and its activities in the static log files, the start time and end time for each user access was identified and converted them into single 30 minutes default sessions.

Static User and Session Identification

Input : $W'[]$ // cleaned log data

Input $W'[]$

For each instance i in $W'[]$

For each field $w(f)$ in $W'[i]$

If(w_i (f)==”IP”)

then

If((w_i (f)=”timestamp”)>=start_time&& w_i (f)<=end_time)

then

Check the IP address and DNS status state.

Use cookies to identify users.

Find the navigation paths between the session time with same session_id.

End If

Else

Session changed;

End if

End for

End for

Limitations of the Static Server Logs:

1. Server log files cannot identify the user's session directly. Since, log files have different types of

session fields, it is difficult to process the session duration and ID in dynamic web log files.

2. A personalized web content is impossible to reconstruct from a large set of web logs.
3. The URL's enclosed in the static web logs contains file names and its request parameters, but no semantic data.Semantic data need to map the URLs to the page navigation and source of request and response to the end user.
4. Dynamic web logs from the web pages (PHP, JSP, ASPX) are difficult to handle user identification and session identification.
5. Static preprocessing methods cannot handle dynamic web packets from client to server.

F) Static Web Pattern Analyzers:

The preprocessed log files are given input to the static analyzers to discover the web usage patterns for decision making. The basic traditional static analyzers such as web frequent mining techniques, tree based decision making techniques and graph based pattern miners have a good pattern discovery performance on the static data compared to dynamic web log data. Frequent pattern mining techniques such as Apriori,FP-growth and high utility miners are used to discover the hidden web usage patterns with minimum support and confidence measures.

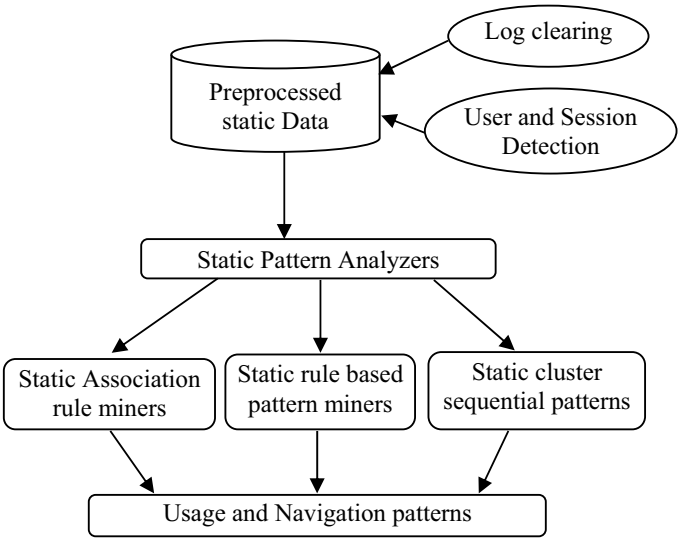

Figure 3.: Different Static Pattern Analyzers

Similarly, static tree based rule mining such as a FP - tree, CP-tree, web access pattern tree[12][13][14][15] and C4.5 are used to generate a large set of frequent patterns with single or multi level usage patterns. Static prediction based models such as Markov model,Bipartile graph, neural network and Bayesian networks[16] are used to predict the sequence of web usage actions and its navigation patterns.

IV. Experimental Results

An experiment was carried out using a log file extracted from the Linux server in three different cases: 1-month, 3-months and 6-months. Initially, there are 500 instances of 1month are chosen in the log file. Static log cleaning of data is done by removing the noisy URL and missing data from a log file.Finally, a cleaned log file is given to different pattern mining techniques for static results.

TABLE 1: Log clustering results

Days	Datasize	Clusters	TP Cluster Rate
30days	#500	6	0.85
60days	#1000	5	0.89
90 days	#2000	7	0.87
120days	#5000	5	0.897

Table 1: Illustrates the graph based cluster model on the static log data of different sizes taken from 30days to 120 days. As the number of instances and clusters increases, the true positive cluster rate and cluster accuracy minimizes and it takes long time to get the specified clusters.

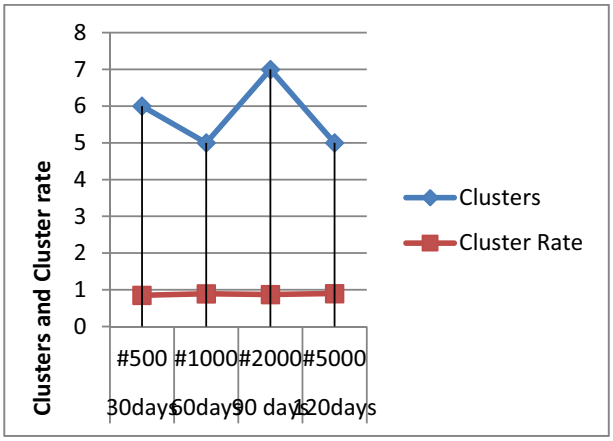

Figure. 4: Number of Clusters and Cluster rate for static logs

Fig. 4, Illustrates the graph based cluster model on the static log data of different sizes taken from 30days to 120 days. As the number of instances and clusters increases, the true positive cluster rate and cluster accuracy minimizes and takes long time to get the specified clusters.

CP-Tree based Association rules

1. responsebytes<= 888.0 AND requestpage != GET /ocart/CLUSTERING?sort=rating&order=ASC&limit=100 HTTP/1.1 -> country != Poland
2. requestbytes<= 88779.0 AND serverres<= 500.0 ->timeduration<= 9994.0
3. timeduration<= 9994.0 AND requestbytes<= 88779.0 -> country != Poland
4. responsebytes>= 251.0 AND requestbytes<= 88779.0 -> country != Poland
5. serverres<= 500.0 AND responsebytes<= 888.0 -> country != Poland
6. requestpage != GET /ocart/CLUSTERING?sort=rating&order=ASC&limit=100 HTTP/1.1 AND serverres<= 500.0 -> country != Poland
7. timeduration<= 9994.0 AND responsebytes<= 888.0 ->serverres<= 500.0
8. requestbytes<= 88779.0 -> country != Poland
9. requestpage != GET /ocart/CLUSTERING?sort=rating&order=ASC&limit=100 HTTP/1.1 AND serverres<= 500.0 -> country != Poland
10. country != Poland AND responsebytes<= 888.0 ->requestpage != GET /ocart/CLUSTERING?sort=rating&order=ASC&limit=100 HTTP/1.1

The evaluation result of CP-Tree model on the static log files is shown below. In this evaluation,the true positive rate along with other statistical measure are computed on the static log files.The classified urls,sessions and its performance are analyzed using computed classification accuracy.

Elapsed time: 19.47s
Number of Iterations :4
F-Measure: 0.92698
Recall : 0.90833
TP rate : 0.94522
FP rate : 0.05478000000000005
Classification Accuracy 0.94506

Static Agglomerative Clustering:

The agglomerative clustering algorithm on the static log file was used to identify the user's url and session wise clustering information. Each instance and its related cluster's information wasshown below.

95.211.135.133,'United States',7:4,681,35542,200,7777,'GET /main/cron/run.php HTTP/1.0'====>C1
157.55.39.29,Brazil,21:32,471,13875,500,6099,'GET /main/cron/run.php HTTP/1.0'====>C3
64.79.100.24,Netherlands,20:18,826,73794,404,5620,'GET /main/cron/run.php HTTP/1.0'====>C1
199.79.62.54,India,10:18,255,39455,404,9916,'GET /results/result.php?no=1313ah04&sem=5 HTTP/1.1'====>C0
117.201.62.143,Canada,14:32,303,29991,500,1050,'GET /results/css/images/side1.jpg HTTP/1.1'====>C3
198.252.44.15,'United States',20:37,522,53759,404,4988,'GET /main/css/chamilo/images/user_icon.png HTTP/1.1'====>C3
175.101.67.26,India,11:4,379,46687,404,2337,'GET/ HTTP/1.1'====>C2
45.55.134.11,'United States',3:35,305,69742,500,5333,'GET /results/css/5grid/core.css HTTP/1.1'====>C3
187.255.73.64,'United States',11:24,884,37683,300,7757,'GET /main/inc/lib/javascript/bootstrap/bootstrap.js HTTP/1.1'====>C1
50.172.216.137,'United States',17:10,858,57969,404,7864,'GET /main/css/chamilo/images/user_password.png HTTP/1.1'====>C1
187.255.73.64,India,11:55,847,68721,200,7903,'GET / HTTP/1.1'====>C2
187.255.73.64,Brazil,14:19,640,5804,300,3313,'GET /main/cron/run.php HTTP/1.0'====>C0
175.101.67.26,China,3:12,338,26128,200,1906,'GET /main/cron/run.php HTTP/1.0'====>C2

Cluster No	Instances(Percentage)
0	120 (24%)
1	179 (36%)
2	59 (12%)
3	142 (28%)

V CONCLUSION

In this paper, a general static model for log cleaning, user & session identification and pattern discovery models are studied on the different log files. This, allows to guarantee an importance of static pattern discovery models on static web log files for decision making. Also, static models only utilize temporal logs for pattern evaluation and navigation, which affects the precision and recall. Experimental results show that the static models on static log data have high performance in terms of F-measure and error rate on the limited data size and limited fields. In future work, we will develop a mathematical model to predict the user's behavior on dynamic web contents based on server log and customer database. Also, dynamic user behavior patterns through the navigation patterns will be analyzed using the user's interest or context information and the existing user's navigation patterns.

REFERENCES

[1] C. J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, S. García,"Web usage mining to improve the design of an e-commerce website: OrOliveSur.com",Expert Systems with Applications: An International Journal , Volume 39 Issue 12.

[2] Yu-Shiang Hung, Kuei-Ling B. Chen, Chi-Ta Yang, Guang-Feng Deng," Web usage mining for analysing elder self-care behavior patterns",Expert Systems with Applications: An International Journal , Volume 40 Issue 2.

[3] Kun Chang Lee, Soonjae Kwon," Online shopping recommendation mechanism and its influence on consumer decisions and behaviors: A causal map approach",Expert

Systems with Applications: An International Journal , Volume 35 Issue 4,2008.
[4] Yao-Te Wang, Anthony J.T. Lee,"Mining Web navigation patterns with a path traversal graph",Expert Systems with Applications: An International Journal , Volume 38 Issue 6.
[5] NisargPathak,ViralShah,ChandramohanAjmeera,"A Memory Efficient Algorithm with Enhance Preprocessing Technique for Web Usage Mining",ICTCS '14 Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies,ACM,2014.
[6] K. SudheerReddy, G. ParthaSaradhiVarma,I. Ramesh Babu,"Preprocessing the web server logs: an illustrative approach for effective usage mining",ACM SIGSOFT Software Engineering Notes archive Volume 37 Issue 3, May 2012.
[7] Maheswari, B.U.; Sumathi, P.," A New Clustering and Preprocessing for Web Log Mining",Computing and Communication Technologies (WCCCT), 2014, 25 - 29, DOI: 10.1109/WCCCT.2014.67.
[8] Vincent S. Tseng , Kawuu Weicheng Lin, Jeng-Chuan Chang," Prediction of user navigation patterns by mining the temporal web usage evolution",January 2008, Volume 12, Issue 2, pp 157-163.
[9] Yan Li,Boqin Feng ,Qinjiao Mao ,"Research on Path Completion Technique in Web Usage Mining",Proceeding ISCSCT '08 Proceedings of the 2008 International Symposium on Computer Science and Computational Technology - Volume 01,Pages 554-559
[10] Baoyao Zhou, Siu Cheung Hui, Alvis C. M. Fong," Discovering and Visualizing Temporal-Based Web Access Behavior",September 2005,WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence
[11] XidongWang,"Discovery of user frequent access patterns on Web usage mining",Computer Supported Cooperative Work in Design, 2004. Proceedings,IEEE.
[12] Gupta, Arora, R. ; Sikarwar, R. ; Saxena, N.," A,Web usage mining using improved Frequent Pattern Tree algorithms"Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 .
[13] Cong-RuiJi; Zhi-Hong Deng," Mining Frequent Ordered Patterns without Candidate Generation",Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007.
[14] Lei Zou; Yansheng Lu; Huaming Zhang; RongHu,"Mining Frequent Induced Subtree Patterns with Subtree-Constraint Data Mining Workshops, 2006. ICDM Workshops 2006.
[15] Liu Jian-ping; Wang Ying; Yang Fan-ding," Incremental Mining Alogorithm Pre-FP in Association Rules Based on FP-tree",Networking and Distributed Computing (ICNDC), 2010
[16] Awad, M.A.; Khalil, I.,"Prediction of User's Web-Browsing Behavior: Application of Markov Model",Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions onYear: 2012, Volume: 42, Issue: 4