Implementation and Analysis of a Hybrid approach for Clustering Weblog.

Chandana S. Khatavkar , Prof. Mangesh Wanjari

Abstract— **Web mining for usage pattern is the key to discover marketing intelligence in e-commerce. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Various web usage mining techniques are used to analyze the user's navigational patterns. Clustering is unsupervised classification of data items into groups called clusters. Recently, algorithms inspired by nature are used for clustering. The study of ant colonies behavior and their self-organizing capabilities is of interest to knowledge retrieval. In this paper, an Ant based clustering algorithm is proposed to discover Web usage patterns (data clusters). This paper includes evaluation and the implementation of Ant based Clustering algorithm. The clustered data is then used to analyze the trends like URL visit Analysis, Traffic, maximum hit and session killing pages with the help of Fuzzy Inference rule.**

Index Terms— **Web usage mining, ant-based clustering, Fuzzy inference rules,**

I. INTRODUCTION

During the past few years the World Wide Web has become the largest and most popular way of communication and information broadcasting. It serves as a platform for exchanging various kinds of information. The volume of information available on the internet is increasing rapidly with the explosive growth of the www. [13]. Web usage mining is the process of knowledge exploitation from the secondary data [3]. Secondary data, mean browser logs, user profiles, web server access logs, registration data, user sessions or transactions, cookies, user queries, mouse clicks and any other data that is the result of interaction with the Web.

Web usage information mining could help to engage new customers, maintain current customers, track customers who are leaving web site, and so on [1]. Usage information can be extracted to increase web server efficiency by prefetching and caching strategies. Based on several researches done in the area of web mining, we can broadly classify it into three domains: web content mining, web structure mining, and web

usage mining. Web content mining is the process of extracting knowledge from the content of the actual web documents (text content, multimedia etc.) Web structure mining is targeting useful knowledge from the web structure, hyperlink references and so on. Web usage mining is the process of extracting useful information from server logs i.e. users history.[8]There are three main tasks for performing Web Usage Mining: Pre-processing, Pattern Discovery and Pattern Analysis.

E-commerce web sites continue to grow in size and complexity, the result of web usage mining with various tools have become critical for a number of application such as web site design , business and marketing decision support and network traffic analysis. So by understanding the interest of user, the companies can establish better customer manager relationship by giving them exactly what they require. Companies can understand the requirements and serve them accordingly[4].They can also increase profitability and productivity based on the profiles generated.

In the following section we give an overview over the related work. Section III explains the proposed methodology and ant-based clustering in more detail. Section IV goes into detail how we implement the proposed method i.e. the experimental procedure of the proposed method and results are shown. Section V shows the uses and significance of the proposed system. We concluded our work in Section VI.

II. RELATED WORK

Web usage mining also known as web log mining, aims to discover interesting and frequent user access patterns from web browsing data stored in the log files of web/proxy servers or browsers. Various data mining techniques such as statistical analysis, association rules, clustering, classifications and sequential pattern mining have been used for mining web usage logs.[10]

[1] proposed a new method to extract navigational patterns from web logs. Ant-based clustering has been used for this purpose. They have numbered every webpage and a sequence

is extracted from the user navigation on the website. Similar sequences are clustered together and then it is displayed in an interpretable format. After the clustering is completed, alignment processing has been applied to the extracted sequences in each cluster and extract the representative for each cluster.

[11] Deals with cluster optimization technique based on fuzzy logic. Clustering technique is used for discovering useful usage patterns. The proposed model uses FCM approach for clustering based on user sessions. The users with similar access patterns are clustered together. Fuzzy Cluster-chase algorithm is used for cluster optimization, to personalize web page clusters of the end users. It is used for eliminating the redundancies occurred in data after clustering done by web usage mining methods.

In [12], the hybrid framework uses an ant colony optimization algorithm to cluster Web usage patterns. This paper proposed an ant clustering algorithm (ACLUSTER) to segregate visitors or find the web usage patterns (data clusters) and a linear genetic programming approach to analyze the visitor trends. The results are compared with the earlier works using self-organizing map and evolutionary fuzzy C-means algorithm to segregate the user access records and several soft computing paradigms to analyze the user access trends.

III. METHODOLOGY

Web usage mining deals with the extraction of efficient usage patterns from web log data, in order to understand and provide the needs of web based applications. The web usage mining process includes the following steps: Data collection, Preprocessing of log file, Pattern discovery based on Ant Colony based algorithm and the pattern analysis with the help of Fuzzy if-then Rules Figure 1.Describes the general frame work for the proposed model.

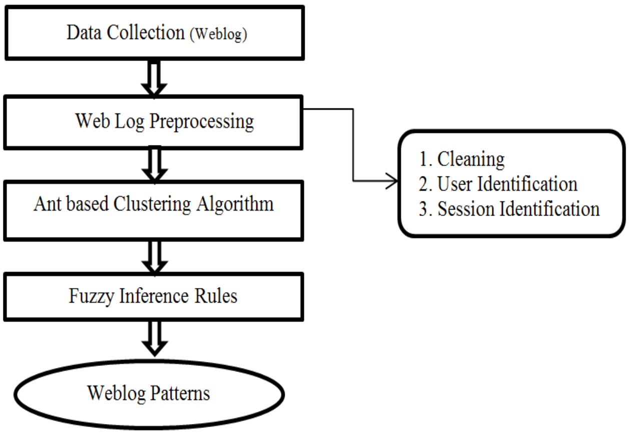
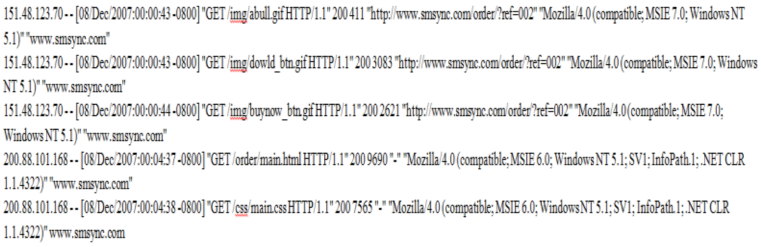


Figure 1. General framework for the proposed model.

I. Data Collection.

The input for the proposed model is web log file. When a user accesses website its navigation activity on the website is recorded in a log file by the web server. Log file is available in two formats. The first is the common log format (CLF) which records the host name and the version of the user's web browser. The second is the extended log format (ELF) The common log format is used in the proposed system. Figure 2 showsthe example of common log data.



```
151.48.123.70 - - [08/Dec/2007:00:00:43 -0800] "GET /img/abull.gif HTTP/1.1" 200 411 "http://www.smsync.com/order/?ref=002" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)" "www.smsync.com"
151.48.123.70 - - [08/Dec/2007:00:00:43 -0800] "GET /img/dowld_btn.gif HTTP/1.1" 200 3083 "http://www.smsync.com/order/?ref=002" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)" "www.smsync.com"
151.48.123.70 - - [08/Dec/2007:00:00:44 -0800] "GET /img/buynow_btn.gif HTTP/1.1" 200 2621 "http://www.smsync.com/order/?ref=002" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)" "www.smsync.com"
200.88.101.168 - - [08/Dec/2007:00:04:37 -0800] "GET /order/main.html HTTP/1.1" 200 9690 "-" "Mozilla/4.0 (compatible;MSIE 6.0; Windows NT 5.1; SV1; InfoPath.1; .NET CLR 1.1.4322)" "www.smsync.com"
200.88.101.168 - - [08/Dec/2007:00:04:38 -0800] "GET /css/main.css HTTP/1.1" 200 7565 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; InfoPath.1; .NET CLR 1.1.4322)" www.smsync.com
```

Figure 2. Example of Common log files record.

II. Preprocessing

Preprocessing is the process of preparing log data for further analysis by removing irrelevant data items. This step can be divided into at least three sub steps: Data Cleaning, User Identification, Session Identification.[1]

A. Data Cleaning

The first step in preprocessing is data cleaning. When a user requests a page, the request is added to the Log File; but if this page contains images, java scripts, flash animations, video, etc., they are added to the Log file as well. Most of the time, these are not needed for pattern discovery and are deleted from log files.

B. User Identification

The second step in preprocessing is User Identification. The fields required for analysis are extracted from the cleaned log file and stored in the database for further processing. There are several methods to identify an user, [1] like detecting cookies, through IP address and user name. [2] Uses cookies to identify users. But they have two main problems: the users can lock the use of cookies, so server cannot store information locally in the user machine; other problem is the user can delete the cookies. Here in the proposed model, IP addresses are considered to identify a particular user.

C. Session Identification

After User Identification, the pages accessed by each user are divided into individual session. A request from a particular user within a predefined time period is considered as a user session. Each user session is identified by session ID. Hence this step is known as Session Identification [5].

II. Pattern Discovery.

As stated in [6], Pattern discovery "draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition". Clustering approach is proposed for pattern discovery. Cluster is a collection of data objects that are similar to one another. In the case of web usage mining the data objects are User sessions generated by preprocessing step. Clustering method is based on Ant-based Clustering algorithm explained in figure 3.

The agents (ants) and data are randomly initialized on a toroidal grid. By moving agents, data is sorted according to its neighbors.Ant can only sense the similarity of the objects in their immediate region.

- Pick(ant is unladen)Pick= Any random sample.
- Drop(ant is laden)Drop= $(\Sigma x/n)*100>=$ threshold.

where x is no of similar objects and n is no of total objects.

Algorithm:

```
1. Start.
2. Spread Data on Grid.
3. Initialize ants.
4. If ant is un-laden then perform following else
move to step 4.3.
4.1 Ant checks for the object.
4.2 Ant picks up object.
4.3 Moves to find next object.
4.4 If object has similarity to another object on grid
then perform following else move to step 4.5
4.4.1 Drop the object next to similar object.
4.4.2 Go to step 4.6.
4.5 Check if end of grid else go to step 4.3.
4.5.1 Drop the object to its original location.
4.6 Check if object is not the last object on the grid.
Else go to step no 5.
4.6.1 Go to step no 4.1.
5. Stop.
```

Figure 3: Ant based Clustering Algorithm.

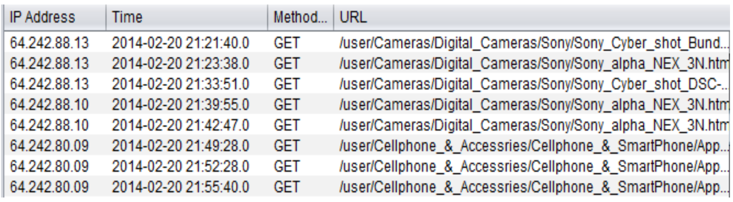
III. Pattern Analysis.

Pattern Analysis is the final stage of WUM (Web Usage Mining), which basically involves extraction of the interesting rules from the output of the pattern discovery process i.e. Clusters. For this Fuzzy If-then Rules are applied to the obtained clusters.

IV. IMPLEMENTATION AND RESULTS

For implementation and analysis phase, the software specifications are Net Beans IDE 8.0.1 and MySql for the database. Web log file i.e input to the system is a e-commerce electronic goods website log, in Common log format. 800 rows of log file is used for performing the web mining process.

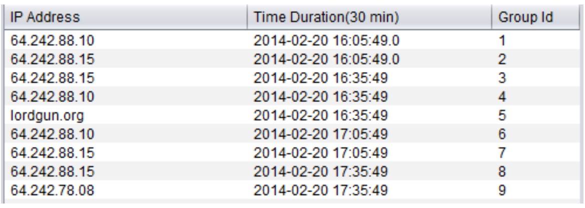
In the Preprocessing phase, we have performed three steps of Preprocessing i.e. Cleaning, User Identification and Session Identification. Data cleaning can be done by checking the suffix of URL name such as gif, jpeg, jpg etc. Required Fields are extracted from the Cleaned Log File and stored in the database for further processing.



IP Address	Time	Method...	URL
64.242.88.13	2014-02-20 21:21:40.0	GET	/user/Cameras/Digital_Cameras/Sony/Sony_Cyber_shot_Bund...
64.242.88.13	2014-02-20 21:23:38.0	GET	/user/Cameras/Digital_Cameras/Sony/Sony_alpha_NEX_3N.htm
64.242.88.13	2014-02-20 21:33:51.0	GET	/user/Cameras/Digital_Cameras/Sony/Sony_Cyber_shot_DSC-...
64.242.88.10	2014-02-20 21:39:55.0	GET	/user/Cameras/Digital_Cameras/Sony/Sony_alpha_NEX_3N.htm
64.242.88.10	2014-02-20 21:42:47.0	GET	/user/Cameras/Digital_Cameras/Sony/Sony_alpha_NEX_3N.htm
64.242.80.09	2014-02-20 21:49:28.0	GET	/user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/App...
64.242.80.09	2014-02-20 21:52:28.0	GET	/user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/App...
64.242.80.09	2014-02-20 21:55:40.0	GET	/user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/App...

Figure 4: Results of User Identification Step.

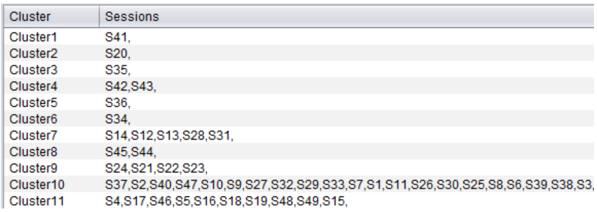
IP addresses are used to identify the users in the user identification step. Required fields are extracted and stored in database as shown in Figure 4.Each request from a particular User within a predefined time period is considered as User Session. Each such user sessions are identified by the session ID. We assume 30 minutes session timeout for the experimental procedure.



IP Address	Time Duration(30 min)	Group Id
64.242.88.10	2014-02-20 16:05:49.0	1
64.242.88.15	2014-02-20 16:05:49.0	2
64.242.88.15	2014-02-20 16:35:49	3
64.242.88.10	2014-02-20 16:35:49	4
lordgun.org	2014-02-20 16:35:49	5
64.242.88.10	2014-02-20 17:05:49	6
64.242.88.15	2014-02-20 17:05:49	7
64.242.88.15	2014-02-20 17:35:49	8
64.242.78.08	2014-02-20 17:35:49	9

Figure 4: Results of Session Identification Step.

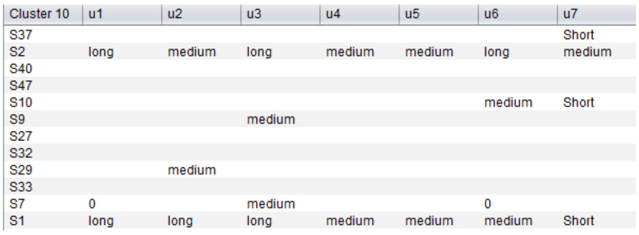
After pre-processing phase, Clustering is performed for Pattern Discovery. For working of proposed Ant based Clustering algorithm [6], sampling of sessions has been done on a toroidal grid [1]. A string of url's, visited by user in a particular session is generated. Sessions are compared using Jaccard's similarity. Figure 5 shows different clusters formed with the sessions.



Cluster	Sessions
Cluster1	S41,
Cluster2	S20,
Cluster3	S35,
Cluster4	S42,S43,
Cluster5	S36,
Cluster6	S34,
Cluster7	S14,S12,S13,S28,S31,
Cluster8	S45,S44,
Cluster9	S24,S21,S22,S23,
Cluster10	S37,S2,S40,S47,S10,S9,S27,S32,S29,S33,S7,S1,S11,S26,S30,S25,S8,S6,S39,S38,S3,
Cluster11	S4,S17,S46,S5,S16,S18,S19,S48,S49,S15,

Figure 5: Result of Ant based Clustering algorithm.

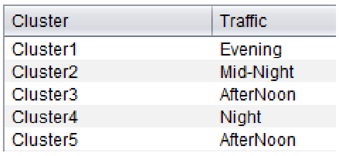
In each session User visits a URL. URL visit analysis is done according to visit time of each user on each URL. It is categorized into short, medium and long. Time spend by user while accessing the website is shown in Figure 6.



Cluster 10	u1	u2	u3	u4	u5	u6	u7
S37							Short
S2	long	medium	long	medium	medium	long	medium
S40							
S47							
S10						medium	Short
S9			medium				
S27							
S32							
S29		medium					
S33							
S7	0		medium			0	
S1	long	long	long	medium	medium	medium	Short

Figure 6: URL Visit Analysis for one of the cluster.

Website traffic is analyzed according to different time ranges. This step helps to identify in which period of day traffic is high. This step will help to attract customers by giving offers when traffic is less. Traffic analysis is done for each cluster.



Cluster	Traffic
Cluster1	Evening
Cluster2	Mid-Night
Cluster3	AfterNoon
Cluster4	Night
Cluster5	AfterNoon

Figure 7: Traffic Analysis.

Certain pages are least interested by users and an assumption is made that they log out their sessions at such pages. These pages are identified and called as session killing pages. Maximum viewed pages and least visited pages are listed so that least visited pages can be refined or more attractive offers can be placed.

As compared to other methods, it has some advantages like it is simple. One only needs to define a suitable similarity function for the clustering step and do not need to involve in complex mathematical relations. Comparison of extensive manual analysis is done with the system results, as shown in below graph.

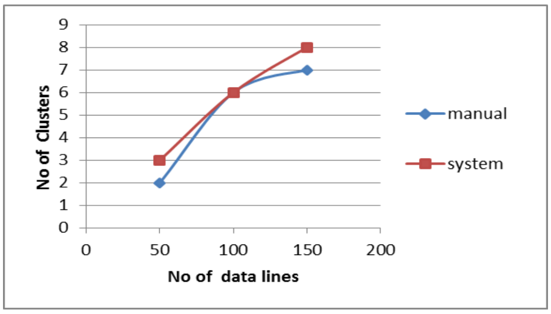


Figure 8 : Result of Ant based Clustering Algorithm on increased data

V. SIGNIFICANCE OF THE PROPOSAL

The proposed method tested tasks like visit time of the user on the website, finding traffic on web page and frequently viewed URLs as suggestion along with session killing pages. These kinds of results are useful for web site owners. They can put their advertisements on maximum hit pages. or to even change the structure of the web site according to users' navigational behavior. The advantage this method is it is simple with less memory utilization and less run time. The listed session killing pages can lead to future enhancement as changing these session killing pages automatically or to redirect the user to other web pages, to increase usage of the website. This system can also be used as base for recommender system.

VI. CONCLUSION AND FUTURE SCOPE.

Preprocessed weblog is used pattern discovery. For pattern discovery clustering is done to segregate the visitors with similar access patterns. The results which were obtained after the analysis were satisfactory and contained valuable information about the Log Files. By analyzing the effect on data using Hybrid approach of Ant colony clustering and fuzzy inference system is that, if there is larger size of data under this algorithm redundancy will increase, so it requires to optimize the data i.e need to remove redundancy of URLs. Many of the people rely on Web sites to access important. information about their interests. To capture the interests of the user, fuzzy inferences are better and hence it makes fuzzy logic most suitable for Web usage mining. The listed session killing pages can lead to future enhancement as changing these session killing pages automatically or to redirect the user to other web pages, to increase usage of the website. This system can also be used as base for recommender system.

REFERENCES

[1] Kobra Etminani Mohammad-R. Akbarzadeh-T. Noorali Raeeji Yanehsari ,"Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", IFSA-EUSFLAT 2009.

[2] M. Eirinaki, M.Vazirgiannis, Web Mining for Web Personalization,Athens University of Economics and Business, 2003.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 1(2), pp. 12-23, 2000.

[4] zahid ansari, a. vinaya babu, waseem ahmed, mohammad fazle azeem ,"a fuzzy set theoretic approach to discover user sessions from web navigational data".

[5] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.

[6] J. Handl, B. Meyer, Improved ant-based clustering and sorting in a document retrieval interface. Proceeding of the Seventh International Conference on Parallel Problem Solving from Nature.

[7] Søren E. Jespersen,Jesper Thorhauge,Torben Bach Pedersen "A Hybrid Approach To Web Usage Mining",

Technical Report 02-5002,Department of Computer Science Aalborg University.

[8] D. Vasumathi & Dr. A Govardhan, "Efficient web usage mining based on formal concept analysis", Journal of Theoretical and Applied Information Technology,2009.

[9] Saroj Bala, S. I. Ahson, R. P. Agarwal ,"An Improved Model for Ant based Clustering", International Journal of Computer Applications (0975 – 8887) Volume 59– No.20, December 2012.

[10] Abhishek Mathur, Trapti Agrawal ,"A Survey: Access Patterns Mining Techniques and ACO", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013.

[11] Nayana Mariya Varghese, Jomina John ,"Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", IEEE 2012.

[12] Ajith abraham, vitorino ramos "web usage mining using artificial ant colony clustering and linear genetic programming"@IEEE 2010.

[13] V. Losarwar, Dr. Madhuri Joshi, "Data Pre-processing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore.

Chandana S. Khatavkar, Mtech Student, Computer Science & Engineering Department,Autonomous, SRCOEM, Nagpur, India.

Prof.Mangesh Wanjari , Asst Professor, Computer Science & Engineering Department,Autonomous, SRCOEM, Nagpur, India.

.