Ant Based Data Reduction In Web Usage Mining Using K-means Clustering Algorithm

Kajal Mengar[1], Prof. Lokesh Gagnani[2], Prof. Dushyantsinh Rathod[3]

[1]M.E. Scholar, [2,3]Assistant Professor
[1,2]Kalol Institute of Technology, [3]Aditya Silver Oak Institute of Technology

***Abstract* - Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs. Data cleaning perform in the Data Preprocessing and Web Usage Mining. The work on data cleaning of web server logs, irrelevant items and useless data can not completely removed and Overlapped data causes difficulty during data retrieving from database. In this paper, we introduce Ant Based Pattern Clustering Algorithm to get pattern data for mining .It also presents Log Cleaner that can filter out much irrelevant, inconsistent data based on the common of their URLs.**

***IndexTerms*— Data Mining, Clustering, Data Reduction, Ant based clustering, Web usage Mining.**

I. INTRODUCTION

Web Mining is technique in data mining to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. In Web Mining, data can be collected at the server side, client-side, proxy servers, or obtained from an organization's database (which contains business data or consolidated Web data). There are many kinds of data that can be used in Web Mining. According to data analysis objective, web mining can be divided into three different types, which are web usage mining, web content mining and web structure mining. Web usage mining is the process of extracting effective information from web server logs.

Clustering analysis plays an important role in data mining field. Data can be grouped into different classes or clusters by clustering analysis. There exists better similarity among the objects in the same class and poorer similarity among the objects in different classes.

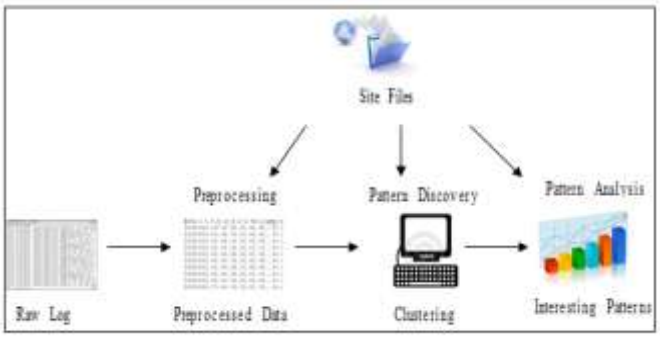

Figure 1: Web usage mining process

II. OBJECTIVE

This paper proposes a clustering method based on Ant Colony Optimization. For clustering ant-based pattern clustering algorithm is applied to pre-processed logs to extract frequent patterns for pattern discovery.

III. TYPES OF LOG FILE FORMAT

Recently, three compositions are available to capture these files:-

1. W3C (World Wide Web Consortium) Extended Log file Format
2. Microsoft IIS (Internet Information Services) Log File
3. NCSA (National Centre for Supercomputing Application) Ordinary Log file Format

All the three are ASCII text layouts. Logging data are recorded in four-digit year format in NCSA and W3C Extended formats. The two digit year format is used in Microsoft IIS log format before 1999 and after that four-digit format is used.

A. W3C Log File Format (World Wide Web Consortium)

W3C Extended log is a customizable ASCII format which has different types of fields. These fields can be parted by spaces. Time can be documented as UTC (Coordinated Universal Time)[7]. Following fields are shown in fig : Client IP-Address, Time-stamp, Method, Protocol Status, URI Stem and Protocol Version.

```
#software: Microsoft Internet Information Services 6.0
#version: 1.0
#Date: 2002-05-02 17:42:15
#Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status cs(User-Agent)
2002-05-02 17:42:15 172.22.255.255 - 172.30.255.255 80 GET /images/picture.jpg - 200
Mozilla/4.0+(  compatible;MSIE+5.5;+Windows+2000+Server)
```

W3C Log File Format [6]

Earlier entry shows that on 02-05-2002 at 05:42 P.M., a user with HTTP version 1.0 and the IP address 172.22.255.255 issued an HTTP GET command for /Default.htm file.

The #Date: field designates when the first most log entry was made and log was created. The #Version: field used to signify the W3C log format. A hyphen (—) shown in the field indicates a placeholder.

B. IIS Log File Format

Microsoft IIS is a non-adjustable ASCII format. This format can record more information than the NCSA format. The IIS format incorporates items like user's IP address, user name, Service status code, request date-time, and number of bytes received. In addition, it includes detailed items like the elapsed time, the number of bytes sent, the action and the target file. Commas are used to split these items which makes format very easy to interpret than the ASCII format, which use spaces for parting. The time is captured as local time. On opening a Microsoft IIS format file in the editor, the entries are seen like the following example in Fig.

```
192.168.114.201, -, 05/15/11, 7:55:20, W3SVC2, SERVER, 172.21.13.45, 4502, 163, 3223, 200, 0, GET,
/index.htm, -,
```

IIS Log File [6]

All the fields are ended with a comma (,). A hyphen(—) works as a placeholder for a certain field which has no valid value.

C. NCSA Log File Format

NCSA Common format is a non-customizable ASCII format which is available for Web sites but not for FTP sites. This captures information about user requests like user name, remote host name, time, date, the number of bytes sent by the server, HTTP status code and request type. Time can be recorded as local time and items can be split by spaces.

```
172.21.13.45 - fred [08/Feb/2010:16:20:14 -0800] "GET /home.htm HTTP/1.0" 200 3401
```

Fig.3 NCSA Log File Format[6]

IV. METHODOLOGY

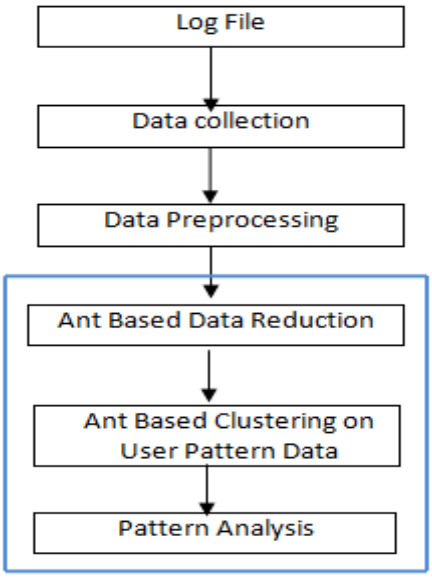

Figure 2: New Architecture

A. Data collection

The input for the web usage mining process is collected from the web log file. During a user session, all navigation activity on the web site is recorded in a log file by the web server. It is a huge repository of web pages and links, accesses web sites are recorded in web logs file. Log file is available in two formats. The first is the common log format and extended log format. '

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"

B. Pre-processing of weblog

By removing irrelevant data items for preparing log data to analysis called as pre processing. Data cleaning is the first step of the process. Cleaning of data can be done by checking the suffix of URL name and deleting the entries such as JPEG, JPG and GIF. The second step in pre-processing is the User Identification. The required fields are extracted from the cleaned log file and stored in the database for further processing. Here, IP addresses are considered to identify a particular user. After data cleaning and User Identification the user sessions are identified. A user session is considered when request of user is within decided time period . Each user session has identified by the session ID **[4].**

C. Ant Based Data Reduction

Basic ant clustering algorithm was proposed by Deneubourg **[3]**. According to this model ants have randomly walk on working area and sense for similarity in nearby objects or not. Based on this information, they would pick the element or drop the element. The probability of picking and dropping an object depends on the objects lying in immediate environment.

Picking probability of an object i :

$$P_{pick}(i) = \left(\frac{k^+}{k^+ + f(i)}\right)^2$$

Dropping probability of an object i is :

$$P_{drop}(i) = \begin{cases} f(i) \; if \; f(i) = k^- \\ 1 \; otherwise \end{cases}$$

where, f =estimation of the fraction of nearby points which is occupied by objects of the same type, and K+ known as constant in this proposed algorithm the data which are going to reduced we just put the flag instead of removing record from data set . so we can identify the performance and accuracy bases on Flag records.

V. ANT BASED PATTERN CLUSTERING ALGORITHM

1. **Input Data Set:** Read ***N*** no of records from clean data source ***FDS***
   For i= 1 to i<=N
   Next
2. For each records ***R*** from data source ***FDS*** find pattern data
3. Read pattern data using specified address from data source ***FDS.***
4. If requested records from frequent data source ***FDS*** with specified pattern then
5. If same record ***R*** from ***FDS = PDS*** then put ***FLAG*** into ***FDS***
6. Make cluster in pattern data source ***PDS***.
7. Else not select those records.
8. End if
9. Next record

Results

The input data is web log file then performing data cleaning to remove unnecessary data items. The cleaned web log is used for pattern discovery. The proposed model uses Ant Colony algorithm for clustering based on user sessions. The users with relevant access patterns will come under the same cluster.

Index_No	Server_IP	Client_IP	URI_Steam	Status_Code	Page_Request	Flag
0	202.71.129.26	10.8.0.15	/Papers/SRSExample-webapp.doc	200	/alldoc.aspx	0
1	202.71.129.26	10.8.0.15	/syllabus.aspx	200	/ss.aspx	0
2	209.85.135.109	10.5.0.54	/starsports.com	200	/cricket.aspx	0
3	59.162.23.130	10.5.0.12	/downloads/index.htm	200	/makemytrip/offer.aspx	0
4	67.218.96.251	10.6.0.20	/downloads/index.htm	200	/admission.aspx	0
5	67.218.96.251	10.6.0.20	/products/W52XXX-series.aspx	200	/product/samsung	0
6	67.218.96.251	10.6.0.20	/it/experienced/index.htm	200	/powerbank	0
7	202.71.129.26	10.8.0.15	http://www.flipkart.com/laptops	200	/sc.aspx	0
8	172.30.255.255	10.5.0.20	http://www.flipkart.com/mobiles	200	/mobiles.html	0
9	209.85.135.109	10.5.0.54	http://www.amazon/Electronics	200	/products.aspx	0
10	67.218.96.251	10.6.0.20	http://in.bookmyshow.com	200	/moviesinfo.aspx	0
11	202.71.129.26	10.8.0.15	/Papers/SRSExample-webapp.doc	200	/alldoc.aspx	1
12	59.162.23.130	10.5.0.12	/downloads/index.htm	200	/makemytrip/offer.aspx	1
13	202.71.129.26	10.8.0.15	/webapp.doc	200	/laptops.aspx	0
14	202.71.129.26	10.8.0.15	/syllabus.aspx	200	/ss.aspx	1
15	209.85.135.109	10.5.0.54	/starsports.com	200	/cricket.aspx	1
16	59.162.23.130	10.5.0.12	/academic/srchpcgn.html	200	/workshop.aspx	0
17	67.218.96.251	10.6.0.20	/downloads/index.htm	200	/admission.aspx	1

Result with Noisy and Flag data

Pass No of Cluster	5	Cluster Cration
Cluster No	202.71.129.26	Create

Index_No	Server_IP	Client_IP
0	202.71.129.26	10.8.0.15
1	202.71.129.26	10.8.0.13
7	202.71.129.26	10.5.0.5
11	202.71.129.26	10.8.0.17
13	202.71.129.26	10.8.0.18
14	202.71.129.26	10.8.0.14
20	202.71.129.26	10.5.0.5
24	202.71.129.26	10.8.0.16
26	202.71.129.26	10.8.0.18
27	202.71.129.26	10.8.0.11
33	202.71.129.26	10.5.0.5
37	202.71.129.26	10.8.0.12
39	202.71.129.26	10.8.0.10
40	202.71.129.26	10.8.0.13
46	202.71.129.26	10.5.0.51
50	202.71.129.26	10.8.0.53

Pattern Cluster 1

Pass No of Cluster	5	Cluster Cration
Cluster No	209.85.135.109	Create

Index_No	Server_IP	Client_IP
2	209.85.135.109	10.5.0.54
9	209.85.135.109	10.6.0.26
15	209.85.135.109	10.5.0.51
22	209.85.135.109	10.6.0.28
28	209.85.135.109	10.5.0.55
35	209.85.135.109	10.6.0.29
41	209.85.135.109	10.5.0.12
48	209.85.135.109	10.6.0.21

Pattern Cluster 2

VI. CONCLUSION

In this paper Log Cleaner filter out approx 60% URL requests with same server IP address which cannot be filtered by traditional data cleaning methods for proxy logs. It create a frequency access data and pattern clustering by implementing pattern clustering techniques to generate pattern cluster for easy access of data from pattern clustering instead of general database. It also improves the future much more accurate and reliable. It gives better performance up to 60% rather than 30% and accuracy compare with old algorithm .

REFERENCES

[1]An Ant-Based Data Reduction Algorithm.
Ismail M. Anwar,Khalid M. Salama,Ashraf M. Abdelbar
[2] A Hybrid approach for Clustering Weblog
Volume 5, Issue 3, March 2015
[3]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
[4]Nayana Mariya Varghese, Jomina John ,"Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic", IEEE 2012.
[5]Shelokar P S, Jayaraman V K, Kulkarni B D. An Ant Colony Approach for Clustering [J]. Analytica Chimica Acta, 2004, 509: 187-195
[6]Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan, Mohamad Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology 48 2008, pp.190-197, DOI: 10.1.1.140.5102
[7]Fang Yuan, Li-Juan Wang, Ge Yu, "Study on Data Pre-processing Algorithm in Web Log Mining", IEEE Nov, 2003, pp.28-32 vol.1, ISBN: 0-7803-8131-9
[8] Nichele C. M. and Becker. K., 2006,"Clustering Web Sessions by Levels of Page Similarity", W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 346 – 350, 2006 © Springer-Verlag Berlin Heidelberg 2006
[9] Renáta Iváncsy, István Vajk, "Frequent Pattern Mining in Web Log Data" , Acta Polytechnica Hungarica, January 2006
[10] Web Usage Mining: A Survey on Preprocessing of Web Log File Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood Department of Computer Science, Muhammad Ali Jinnah University, Islamabad, Pakistan
[11] Alphy, S.Prabakaran, "Cluster Optimization for Improved web Usage Mining using Ant- Nestmate Approach", IEEE-International Conference on Recent Trends in Information Technology, June 3-5, 2011
[12] Log files formats", http://www.w3c.org, Access Date: [5th of Dec, 2012-10 PM].
[13] 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.
[14] Banerjee, A. and J. Ghosh (2001). Clickstream Clustering Chicago (2001)
[15] mrs. v. sujatha, dr. punithaval li ," an approach to user navigation pattern based on ant based clustering and classification using decision trees", 2010.