Web Data Mining using Pattern Matching and Two Phase Algorithm

Siddhesh Gore[#1], Pavan Bawdane[#2], Atul Zore[#3], Viral Parmar[#4], Jyothi Arun*[5]

[1]sidh2106@gmail.com,[2]bawdane3@gmail.com,[3]zoreatul1623@gmail.com,[4]viralparmar099@gmail.com

*Assistant Professor, [5]amritajyothi87@gmail.com

Department of Information Technology, Atharva College of Engineering, Mumbai University
Maharashtra, India

***Abstract* — Due to the advancement in computer technologies to a higher extent, the information available for every context is huge. This has paved a way for increasing e-commerce users to extract useful information from the World Wide Web. Web Mining is divided into Web Content Mining, Web Structuring Mining, and Web Usage Mining. The user access patterns are saved as server access logs in Web Servers and useful results are being obtained. Pattern matching method is used for the cleaning and sorting purpose. The 2 phase algorithm is being used in this to obtain effective results which are used under data mining domain. The web logs obtained from various web servers have become an important data source for data mining and machine learning. In the pre-processing phase the data is being cleaned, analyzed and sent forward for further utility mining.**

***Keywords*— two phase algorithm, pattern matching method, web log, user access pattern, cleaning, web usage mining.**

I. Introduction

The important work of the project is solely based on the web log files or web data acquisition. The secondary importance gets to personalization, caching etc. The results of the web usage mining are used for decision management i.e. it helps for analysis, satisfaction of customer, to make important decisions about the enterprise. To identify and study the users' browsing behaviours are major issues. Predicting the users' requirement can be made by looking at the previous browsing history of user. By implementing the prediction, it provides a growth in market, trends, promotion, product supply etc. It basically helps to make an enterprise achieve growth at a faster rate. Earlier, many technologies were acted upon like Lee and Fu proposed Two Levels of Prediction Model, but it decreased the prediction scale.

Our approach towards this is to mark an important aspect by collecting huge data logs from World Wide Web. To create useful access patterns from them. The web log files contain information which has IP address, browser name, session time, country/city from which it is accessed, date, time. This data will prove important to provide good statistical information. Two Phase Algorithm cleans the raw data and sorts it accordingly, and when it has to be accessed it can be done by getting on the particular attribute pointing the array index in which it is stored.

II. Problem Definition

The normal procedure of data pre-processing includes 5 steps: data cleaning, user identification, and path completion and user transaction. When data pre-processing methods were merged with web-log data, various challenges were faced. As web pages contain more and more advertisements which contain lots of noise which in turn affects data mining [1] and also lacks information. Earlier data which was collected was not extracted by pattern format which in turn gives clear picture of statistics.

We are proposing a new technology for data cleaning and sorting with the help of pattern matching method. Additionally, the Two Phase Algorithm is being used for data pre-processing. Whenever the user visits a particular web page, all the data is collected in the .txt called web-log. For cleaning web-log is collected is collected in .txt file which is in improper format. Cleaning will result in converting web-log data from .txt file into .csv file. Web data mining algorithm helps in extracting the patterns from users' behaviour which helps in making various decisions. Earlier, two-phase algorithm was used in Rubik's cube. But, now it will be used in processing web-data.

III. Existing System

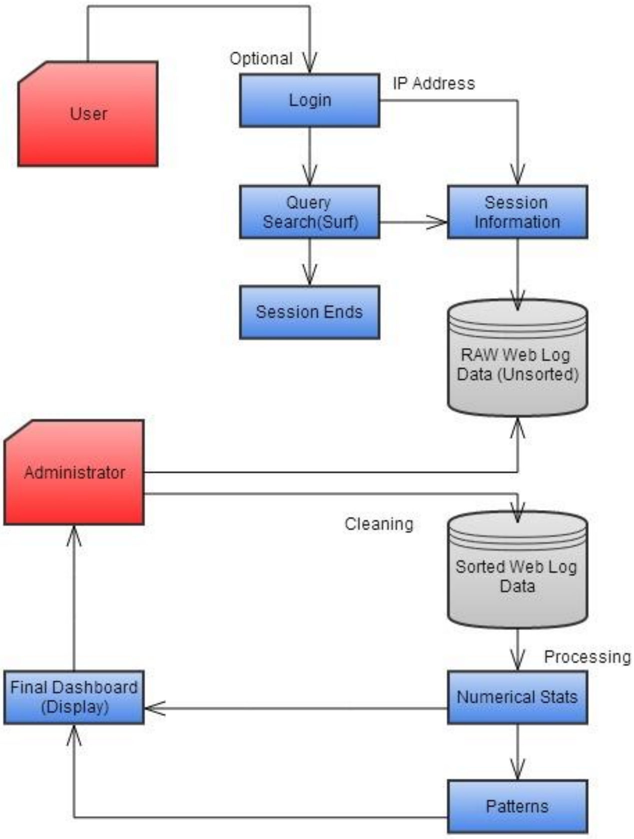
In the provided system, data pre-processing method was used for this purpose. The normal procedure of data pre-processing includes 5 steps: data cleaning, user identification, user session identification, and path completion and user transaction identification. While applying these to the enterprise proxy log, we encounter some new challenges. Web pages are becoming more and more colourful with attachments such as advertisements. It makes the normal data cleaning methods still have a lot of noisy pages which affects the data mining. Besides, this also disables the step of path completion for lacking information.

A. Other Technologies

S. Madria [2] proposed detailed methodology on discovering interesting patterns and facts with the help of connectivity in the web subset and collection of connected web documents.

Tasawar [3] put forth a pre-processing methodology under the web usage mining using hierarchical clustering. This technology enabled data pre-processing on large scale and helped convert categorical web log data into numerical data. Yaxiu [4] proposed fuzzy clustering which involves two important factors i.e., page clicks and web browsing time which are saved in web log data. Houqun [5] put forth multipath segmentation which includes segments and clusters. It enhances the user efficiency by accessing the user access path. Jianxi [6] proposed web usage mining based on the fuzzy clustering.

IV. Architecture



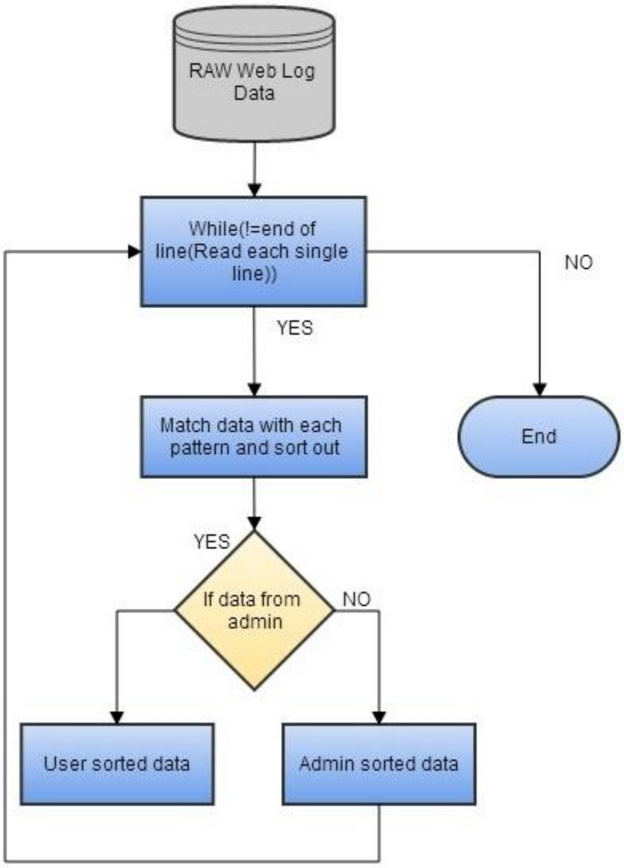
A. User

Whenever the user starts surfing the web/internet for some queries, a session is created. An IP address [7] from the session information is used to authenticate the user. Therefore, when a user visits certain page, a web log is created which comprises of fields like IP Address, Date, Timestamp, GET/POST Request, Host Address, Browser and Operating System. This web log data which we get from session information is initially raw and in unsorted manner. This raw web log data is passed to server/admin.

B. Admin

Admin is responsible for activities such as cleaning the raw web-data to remove redundant data. After cleaning the raw web data it is converted into sorted web log data which is in proper format useful for further processing. A pattern matching method comprised of regular expressions is used for sorting and cleaning the raw web log data. It is useful for comparing and matching the syntax of string characters. After processing sorted web log data we make use of two-phase algorithm to get numerical statistics and patters of users' behaviour. Numerical statistics and patter are displayed in a user friendly, graphical interface which is visible to administrator only.

C. Data Cleaning

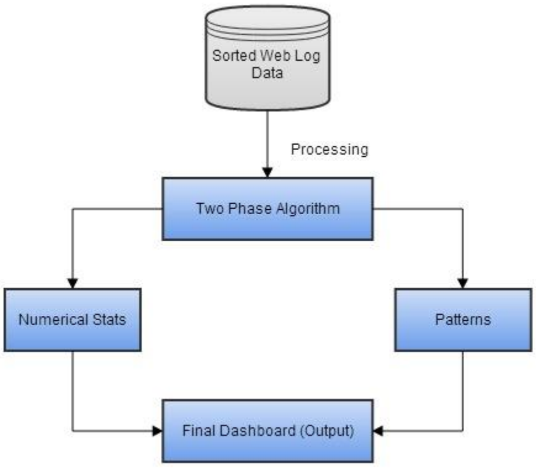

RAW Web Log Data
While(!=end of line(Read each single line))
NO
YES
Match data with each pattern and sort out
End
YES
If data from admin
NO
User sorted data
Admin sorted data

In data cleaning process, a RAW web log data is passed through the while loop. This while loop reads each single line and compares the data with each pattern according to the syntax. If there is no more data available, it goes into the end state. At the end, user's log data is kept separately from the administrator's sorted data.

D. Pattern Matching

In pattern matching phase, each string from the log data is compared with the regular expressions in the php code. If the pattern from the regular expression matches up with the string from the web log data, it is counted as the valid entry. It is then considered useful for further processing.

E. *Data Processing*



In data processing phase, a sorted data retrieved from the data cleaning and pattern matching process is used. A Two Phase Algorithm [8] is applied on this data to yield numerical statistics and patterns. These two entities are used to show the final output to the administrator.

V. ALGORITHM

A. *Data Cleaning Algorithm*

```
Step1: Read LogRecord from Web Server Log File
Step2: If(LogRecord.url-stem(gif.jpegjpg.cssjs)) AND
(LogRecord.method=’GET’) AND
(LogRecord.Sc-status<>(301,404,500)AND
(LogRecord.Useragent<>Crawler.Spider.Robot))
Then
Match(“Pattern match” logRecord, (ip, u_id, u_name, date, time, browser, request, url, protocol, status, user-agent))
insert above match data from LogRecord into LogDatabase(.csv, .txt, .log)
End of If condition.
Step3: Repeat the above two steps until eof(Web Server Log File)
Step4: Stop the process.
```

B. *Two Phase Algorithm*

Input –

- Set of a items z = $\{i_1, i_2 \ldots i_a\}$ with each I_g and Profit value of P_g, where g = 1 to a;
- Transaction Database - D = $\{T_1, T_2, T_3 \ldots T_n\}$ where a subset of items are included in each transaction.
- The minimum threshold variable th.

Output

Set of Utility itemsets

Step 1 - Calculate utility value U_{gk} for each item I_g in each transaction T_k, therefore

$$U_{gk} = Q_{gk} * P_g$$

Where
Q_{gk} = Number of I_g in T_k for g=1 to a and k = 1 to n.

Step 2 - Finding maximum utility value i.e., MU_k for each transaction T_k

MU_k = max $\{U_{1k}, U_{2k} \ldots\ldots U_{ak}\}$ where k=1 to n

Step 3 - Calculating upper bound for utility value i.e., UB_g for each item I_g.

Therefore, summation of all the maximal utilities for each transactions including is given by –

$$ub_g = \sum_{i_g \in Tk} au_k$$

Step 4 - Checking of upper bound value of utility of an item I_g with threshold value th. If I_g is greater than or equal to th (threshold) value then put it in the set of candidate utility 1-itemsets i.e., C_1

$$C_1 = \{i_g \mid ub_g \geq th, 1 \leq g \leq a\}$$

Step 5 - Set z = 1 where z represents the number of items in current candidate utility itemsets.

Step 6 - Generating candidate set C_{z+1} from C_z with all the z-sub-itemsets in each candidate C_{z+1} which is contained in C_z.

Step 7 - Calculating upper bound utility UBs of each candidate utility z+1 itemsets where the summation of maximal utilities of transations which includes s

$$ub_s = \sum_{s \in T_k} mu_k$$

Step 8 - Checking average utility of upper bound of each candidate z+1 itemsets s is greater than or equal to th (threshold). Now, if s fails to satisfy above condition, eliminate it from C_{z+1}.

New C_{z+1} = {s | $ub_s \geq th, s \in original\ C_{z+1}\}$

Step 9 - If C_{z+1} = null, do the next step else
set z = z + 1 and repeat steps 6 to 9.

Step 10 - For each candidate average utility itemset s, find actual average utility avg_s as

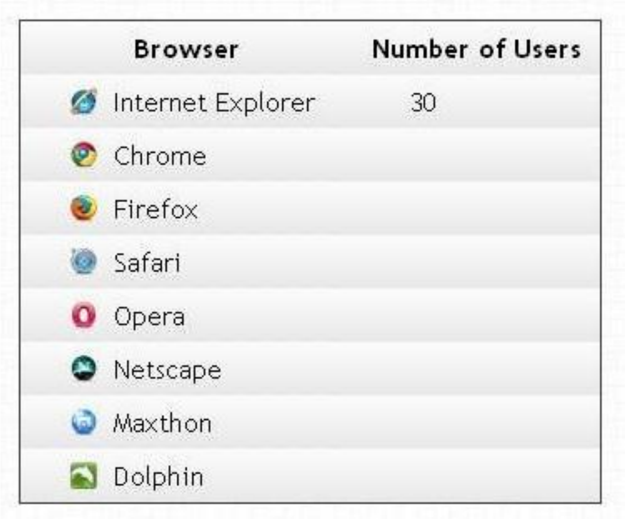
$$avg_s = {\sum_{s \in Tk} \sum_{i_g \in s} ug_k} / {|s|}$$

Where
u_{gk} = utility value of each item ij in transaction T_k
|s| = Number of items in s

Step 11 = checking whether actual average utility value i.e., avg_s for each candidate average utility item set s is greater than or equal to threshold th.

If s satisfies the above mentioned condition, put it in the set of high average utility item sets, HAU.

VI. EXPERIMENTAL RESULTS



Browser	Number of Users
Internet Explorer	30
Chrome	
Firefox	
Safari	
Opera	
Netscape	
Maxthon	
Dolphin	

The administrator dashboard consist results like the image shown above. It consists of well formatted data such as different types of internet browsers, number of users' using particular browser, and type of request made (GET/POST). Additionally it includes visits according to the dates, time and operating system this is used to access the site.

VII. CONCLUSIONS

Web Utility Mining applications run on the fundamental concepts of data pre-processing. It is necessary to process the data before send it out for various mining techniques and pattern generation. This data can further be used for predicting user's behaviour. In this paper, we have proposed a pattern matching method using regular expressions for data pre-processing. We have also proposed a two phase algorithm which can be used for removing redundant information and generating patterns.

In future, this can be used to track the live feed/data of the user's visits to the website. This will further enhance the analytical skills and help site owners to show only the relevant information to the respective users.

REFERENCES

[1] Swapna Mallipeddi, D.N.V.S.L.S.Indira - High Utility Mining Algorithm for Pre-processed web data – International Journal of Computer Trends and Technology Volume 3 Issue 3 – 2012.

[2] S. Madria - Research Issues in Web Data Mining.

[3] Jianxi Zhang, Peiying Zhao, Lin Shang and Lunsheng Wang, "Web usage mining based on fuzzy clustering in identifying target group", ISECS International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009.

[4] Houqun Yang, Jingsheng Lei and Fa Fu, "An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 4, Pp. 644-648, 2007.

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan- "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", 2008.

[6] Bamshad Mobasher, "Data Mining for Web Personalization," LCNS, Springer-Verleg Berlin Heidelberg, 2007.

[7] Mark E. Snyder, Ravi Sundaram, Mayur Thakur- "Preprocessing DNS Log Data for Effective Data Mining", 2008.

[8] Two Phase Utility Mining Algorithm G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi.