PREPROCESSING OF WEB LOG DATA FOR WEB PERSONALIZATION

Rekha Sundari.M[1], Srininvas.Y[1], Prasad Reddy.PVGD[3]
[1]GITAM University, GITAM Institute of Technology,
Rushikonds, Visakhapatnam, Andhra Pradesh, India
[3]Andhra University, Department of CS&SE,
Visakhapatnam, Andhra Pradesh, India

ABSTRACT

World Wide Web has its impact on almost every facet of human lives. It is the prevailing and most popularly known information source that is very simply assessable and searchable. Prior to web, useful information can be gathered by referring a document or gathering the data from the expert's in the related areas, but with the rapid advancements in information technology, web has rigorously changed data seeking behavior of the users. The rapid growth in size and use of World Wide Web with its unique characteristics made web data mining an upcoming and area of demand in the present era. Web Usage Mining is the process of finding information from web usage logs that contain the click behavior of the user to attain the information the user needs. The procedure that is performed to improve the quality of raw data so as to improve the efficiency and ease of implementing pattern discovery methods is called data preprocessing. The web usage data that is taken from the server logs has to undergo many preprocessing phases so as to transform the data into a form that can be applicable to required mining task. In this paper we elucidate different preprocessing techniques applied to three different data sets.

KEYWORDS: *Pattern discovery, Sequential patterns, Feature Reduction.*

I. INTRODUCTION

Web usage mining (WUM) plays a vital role in analyzing navigational behavior of a visitor to the web site. With the explosive growth of E-marketing this application of WUM helps business analysts to enhance their business relationships with the customers by analyzing their click stream behavior in general and concentrating on their special interests in particular (DoruTanasa et al 2004).As the web and its usage continues to grow, so grows the need to analyze web data and extract all manner of useful data from it. By discovering different access patterns of different users web designers can easily improve the link structure of the web sites as well as improve web server performance and provide good service for the customers (Xidong Wang et al 2003).The web usage data that is available in web server logs cannot be applied to any pattern discovery algorithms without preprocessing, as they contain heterogeneous data from multiple data sources. Web log data that serves as primary input for the WUM process has typical characteristics like high volume, high dimensionality and high sparsity. These characteristics of web log data enforce problems like computational complexity and inability of mining algorithms to extract interesting measures from the web log data (Tahirahasan et al 2009). This data might be an image, audio, video or an advertisement from other websites. Pre-processing of Web data to make it suitable for mining was identified as one of the key issues for Web mining (Cooley, Mobasher & Srivastava1999). The paper is organized as follows: section-II highlights the related work in the area of web usage mining, section-III explains the need for preprocessing, section-IV presents the different datasets, their formats, and the procedures applied on the datasets to acquire the data for suitable application of sequential pattern mining algorithms, section-V gives the results description and section-VI summarizes the paper with conclusion and future scope.

II. RELATED WORK

Hussain et al (2010) proposed a framework for web session clustering at preprocessing level of web usage mining. The framework covered the data preprocessing steps to prepare the web log data and convert the categorical web log data into numerical data. A session vector is obtained, so that appropriate similarity and swarm optimization could be applied to cluster the web log data. The hierarchical cluster based approach will enhance the existing web session techniques for more structured information about the user sessions. For clustering phase, a merger of Particle swarm optimization (PSO) and agglomerative algorithms was applied. In first half of clustering algorithm, the PSO algorithm was implemented and obtained the set of winning sessions based on “Angular Separation” and “Canberra Distance”. In second half, the agglomerative algorithm, for hierarchical sessionization of web session was applied. **Kanchana. P and M. Punithavalli (2011)** proposed a hierarchical cluster based preprocessing methodology for web usage mining and a Fuzzy possibilistic clustering algorithm to handle heterozinity of users behavior. In this process the users browsing pattern is forecasted and is utilized for personalization. The two level of prediction model using the markov model and Bayesian theorem is modified by the authors concentrating on preprocessing of the data. The user data is divided as training data and test data, the training data is given as input to the hierarchical clustering algorithm is used for classifying the users browsing behavior and in next step testing data will be passed to the prediction model.

III. NEED FOR PREPROCESSING

A click stream is a sequence of pages visited by a user through a web site in a particular period of time. Each hit against the server corresponding to a http request generates an entry in server access log files. Each log entry contains the date and time of request, the IP address of the client, the content requested, status of the request, the method used and many more things. Each request is logged separately, so at the stage of preprocessing all of these requests must be aggregated in order to structure the data so that it gives some information (Faccaet al.2005).

2.1 Steps Involved In Pre-Processing

2.1.1 Data Cleaning

Data Cleaning is the procedure of removing all unwanted and irrelevant data that does not play any role in our analysis. This irrelevant data may be image, audio or a video file. These files can be removed based on the application of WUM. This process also removes requests made by web robots that simplifies the mining task by removing uninteresting sessions from the log file.

2.1.2 User Identification

The web log file contains the computers name and the user login for web sites requiring user registration. This information is used for user identification. Each user is uniquely represented by the pages visited.

2.1.3 Page Identification

Identifies unique structure of the web site and the navigation links between them.

2.1.4 Session Identification

For each time the user visited the web site, session identification determines the pages requested, the order of the requests and the time spent in each page.

2.1.5 Path Completion

Acquire knowledge of site topology that is necessary to complete the paths.

Substantial preprocessing of click stream data is required before analyzing user browsing behavior. The data analyzed after preprocessing is integrated and transformed in to understandable structure that can be further applied to pattern discovery algorithms. The different stages in data preprocessing are presented in the following figure 1.

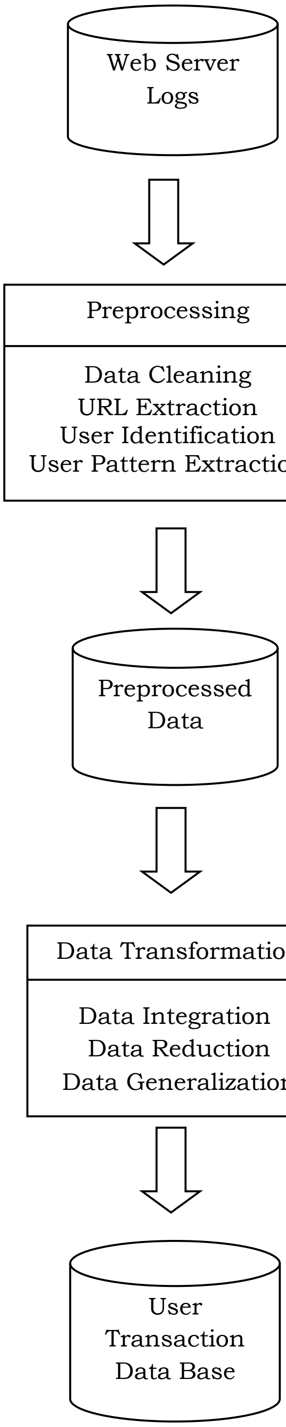


Figure 1: Web Usage Mining Process

After preprocessing of click stream data, the analyst can be able to analyze things like:

1. The page that is frequently/not frequently viewed by many users
2. The page that is not frequently viewed by many users
3. The order of the pages viewed

4. The group of pages the user is interested in
5. The web page that is used as the entry page for most of the users
6. The web page that is exit page for most of the users
7. The time the visitor spent on the web site
8. The special interests of the visitor
9. The visitor future expectations and soon.

In this paper we have considered three datasets to elucidate some of these preprocessing tasks: The msnbc dataset (Pallis, George et al 2007) the cti dataset and other student data set from GITAM University web log data.

IV. DATASETS

3.1 MSNBC DATA SET

3.1.1 Data Set Description

The data comes from Internet information logs of msnbc.com and news relation portion of msn.com for a single day. Each line indicates the page views of a single user for that particular day, the server recorded requests of the user at the level of page category but not at the URL level, the page requests served through caching mechanism were not recorded in the server and hence there present in the data cannot be visualized.

The different page categories are frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports".

Each page category is associated with a number like front page as 1, news as 2 and so on respectively. Each row shown above indicates the hits of a single user in that 24 hour period. For example in the sample sequences given above user 1 hits page 1 that is “frontpage” twice and user 2 hits page 2 that is “news” once. The sample sequences in the original format of the data set are represented in Table 1.The samples of sequences are:

1, 1
2
3, 2, 4, 2, 3, 3
2, 9, 9, 12
1,2,11,15,8
1, 12, 12, 8

Table 1: Representing the basic format of the data set

3.1.2 Preprocessing

We used visits (user, page category) to denote the frequency of a user who browses Web page category of the Web site during a period of time. Suppose frontpage is a Web page which has been visited by user Userl 2-times and by user User2 5-times during a given period of time, the state of the Web page category frontpage is

State(frontpage)= {frontpage, {Userl, 2},{User2, 5}}.

After processing the data as explained, analysis can be made in two ways, user centric analysis and data centric analysis. The first method analyses the users by grouping the users with identical browsing patterns, the second method group’s pages with their highest number of hits. The selection of the methodology depends upon the application of the web miner. As our main goal is to personalize the user browsing behavior, the first methodology is followed by making the following assumptions

1. The users with same interests should have the similar navigational behavior
2. Related Web pages should be navigated by the users with same interests.

3. The general navigational behavior is not changeable during a given period of time for a given user, although different users' browsing patterns maybe different during the specified period of time.

Based on the above assumptions, we can draw user clusters from Web logs by the analysis of users' browsing information during the period of time. Table 2 represents the page visits of each user in integers and Table 3 Represents the same with the names of the pages.

Table 2: Sample user sequences represented as integers

User	Sequence
1	1, 1
2	2
3	3, 2, 4, 2, 3, 3
4	2, 9, 9, 12
5	1, 2, 11, 15, 8
6	1, 12, 12, 8

Table 3: A sample of user sequences represented with page categories

User	Sequence
1	frontpage→frontpage
2	News
3	tech→news→local→news→tech→tech
4	news→health→health→sports
5	frontpage→news→business→travel→weather
6	frontpage→sports→sports→weather

3.1.3 Vector matrix representation:

Before applying usage data for analysis to any clustering algorithm, we constructed vector matrix for user and page category. The user-page category matrix M is used to describe the relationship between user and web pages the user navigated. The entries in the matrix M correspond to the "count"-the number of times user visited the page. Let p be number of pages and u be number of users the matrix M can be represented as M_{uxp} and the Table 4 represents the same in table form with the user and page categories as rows and columns the entry value designates the number of visits to each page.

$$M_{uxp} = \begin{pmatrix} \text{visits}(1,1) & \text{visits}(1,2)\ldots\ldots & \text{visits}(1,j)\ldots. & \text{visits}(1,p) \\ \text{visits}(2,1) & \text{visits}(2,2)\ldots. & \text{visits}(2,j)\ldots. & \text{visits}(2,p) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \text{visits}(i,1) & \text{visits}(i,2)\ldots\ldots & \text{visits}(i,j)\ldots.. & \text{visits}(i,p) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \text{visits}(u,1) & \text{visits}(u,2)\ldots\ldots & \text{visits}(u,j)\ldots. & \text{visits}(u,p) \end{pmatrix}$$

Table 4: Sample user sequences represented by their No. of visits to a Page

User	Frontpage	news	Tech	Local..
1	2	0	0	0
2	1	0	0	0
3	0	2	3	1
4	0	1	0	0
5	1	1	0	0
6	1	0	0	0

3.2 CTI DATA SET

3.2.1 Data Set Description

The dataset contains the preprocessed and permeate sessionized data of Depaul university CTI web server, for a two week period .The preprocessed data contains 13745 sessions and 683 page views or URL's. The data set contains many files representing different data about user navigational behavior. We illustrate below, the data format of only those files that we used in our application.
Cti.cod consists of the unique page views of the website visited by the users. The format of the file is represented in Table 5 where each page view name is assigned a unique ID.

Table 5: Cti.cod contains the different page views along with their id's.

URL Name	URL id
URL 0	0
URL 1	1
URL 2	2
...	…
URL 682	682

Cti.tra file contains data about the sequence of pages visited in a single session. These page views are not ordered according to their page view id's but according to their sequence of visits.

Table 6: Cti.tra representing sequence of page visits in a single session

Session ID	List of URL's visited		
Session 1	URL i	URL j	URL x
Session 2	URL z	URL y	URL k
….	…	….	….
Sessionn	URL a	URL c	URL b

Cti.std file contains data representing a session-page view matrix where each row represents a session and each column represents a page view. The value in the column indicates the amount of time spent on each page view during the session time(session, page view). Suppose if S is the number of sessions and p is the number of page views the vector matrix is represented as M_{sxp}, and the entries are the time spent in each page view. The data is represented in the Table 7.The duration spent in page view is maxed at 999 seconds.

$$\begin{pmatrix} \text{time}(1,1) & \text{time}(1,2)\ldots\ldots & \text{time}(1,j)\ldots. & \text{time}(1,p) \\ \text{time}(2,1) & \text{time}(2,2)\ldots. & \text{time}(2,j)\ldots. & \text{time}(2,p) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \text{time}(i,1) & \text{time}(i,2)\ldots\ldots & \text{time}(i,j)\ldots.. & \text{time}(i,p) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \text{time}(s,1) & \text{time}(u,2)\ldots\ldots & \text{time}(u,j)\ldots. & \text{time}(s,p) \end{pmatrix}$$

Table 7: Cti.std representing session-page view matrix

	URL 0	URL 1	URL 2…….	URL 682
Session 1	T	T	0	T
Session 2	T	T	T	T
Session 3	T	T	T	T
………	….	…	…	…
Session 13745	T8	T2	T1	$T_{max=999}$

Cti.nav file contains data about the session number and user id where multiple consecutive sessions can be associated to the same user id. Table 8 represents the data format of Cti.nav. Under each of this data it again contains three fields, time stamp, page view accessed and the referrer, as we can have the data about the time spent and page view accessed from the cti.cod,cti.tracti.std ,in our application referrer site is not considered, only extraction of session id and user id is performed.

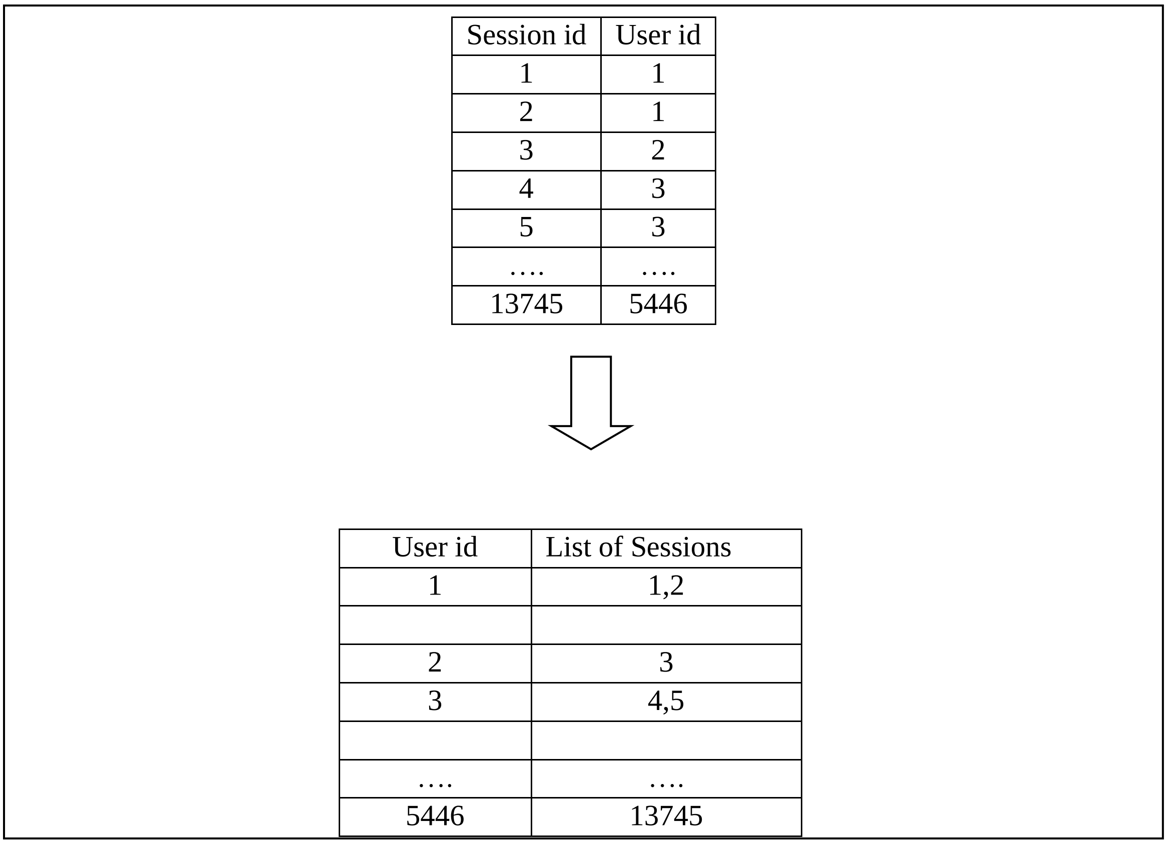
Table 8: Cti.nav representing number of sessions participated by each user.

Session id	User id
1	1
2	1
3	2
4	3
5	3
….	….
13745	5446

3.2.2 Preprocessing:

We used the file cti.cod to identify different page views that are present in the data set as they represent fields in the final matrix we construct. Cti.std is used to extract the time spent in each page in each session and cti.nav is used to identify the sessions each user is participated in. In this data set a single user is participating in multiple sessions .To identify all the page views the user navigated in all the sessions first we have to identify the number of sessions each user is involved in and then combine all the corresponding sessions rows in cti.std into one row that represents the time spent by the user in all the page views during all the sessions using the following steps. Figure 2 represents the way how multiple sessions participated by a user are identified.

Step 1: Identifying multiple sessions that are participated by a single user

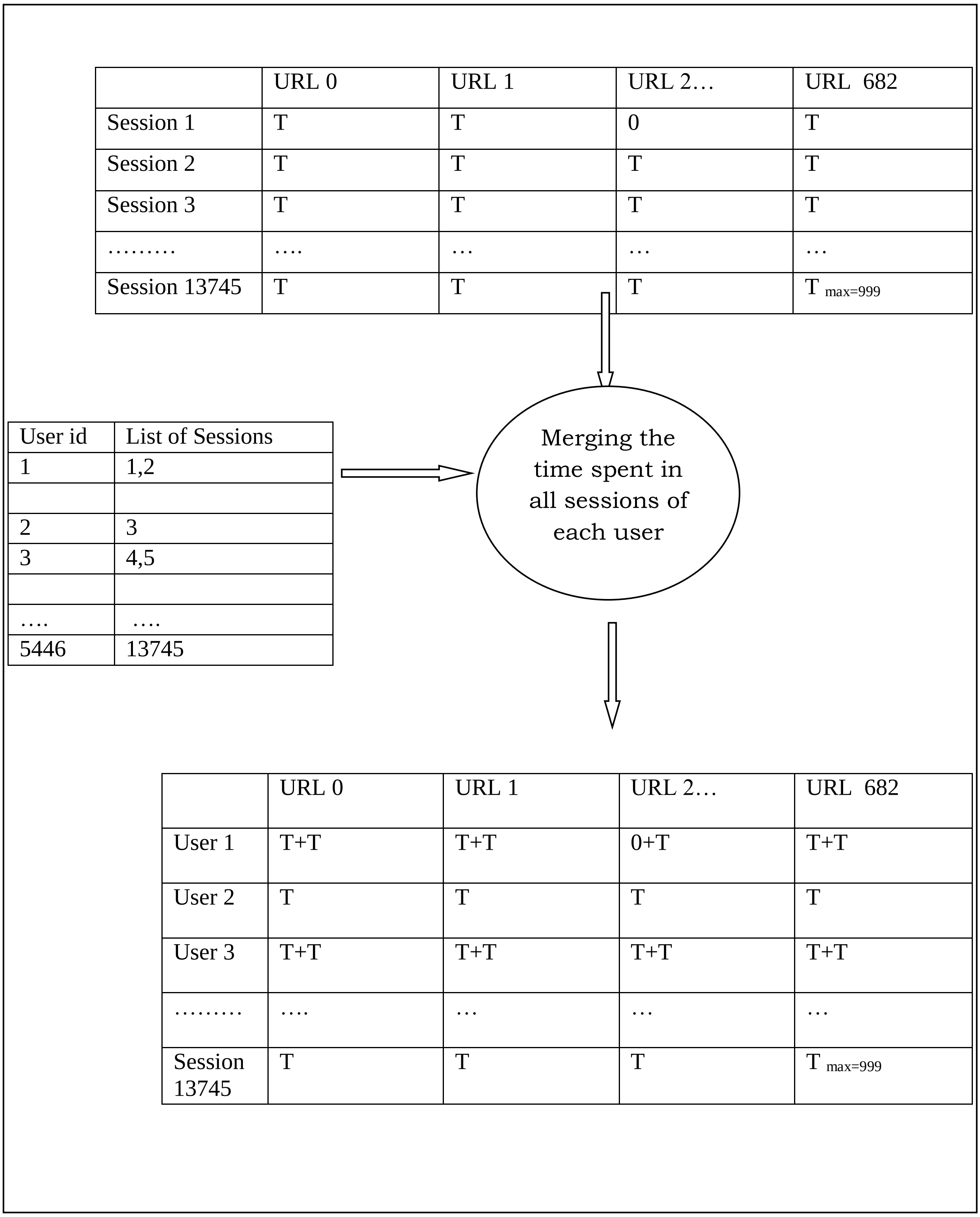


Session id	User id
1	1
2	1
3	2
4	3
5	3
….	….
13745	5446

User id	List of Sessions
1	1,2
2	3
3	4,5
….	….
5446	13745

Figure 2: Process of Identifying Multiple Sessions of a Single user

Step 2: Combining the identified multiple session rows of each user from session versus page view matrix M_{sxp} of cti.std into a single row, to form user versus page view matrix that represents the time spent by each user in each page view in all sessions. The process of deriving the resultant matrix from other two inputs available is explained in figure 3.



	URL 0	URL 1	URL 2…	URL 682
Session 1	T	T	0	T
Session 2	T	T	T	T
Session 3	T	T	T	T
………	….	…	…	…
Session 13745	T	T	T	$T_{max=999}$

User id	List of Sessions
1	1,2
2	3
3	4,5
….	….
5446	13745

	URL 0	URL 1	URL 2…	URL 682
User 1	T+T	T+T	0+T	T+T
User 2	T	T	T	T
User 3	T+T	T+T	T+T	T+T
………	….	…	…	…
Session 13745	T	T	T	$T_{max=999}$

Figure 3: User versus Page view matrix

3.2.3 Feature Reduction:

The main goal in our study is personalization, which mainly concentrates on user navigational behavior. In order to concentrate on mostly visited pages, the number of users visited each page is to be calculated and all the pages that are not visited are to be identified. In the preprocessing stage , we identified that for each column of the user – page view matrix atleast 15 users are not visiting each page, hence we deleted the pages with low visit rate, the number of visits made to each page is calculated by applying the functions ifvalue() and count() on columns, which returns 1 if the entry is nonzero value and returns zero if the entry is zero and count returns the total number of one's respectively. The total number of one's gives the number of visits to each page

Table 9: Final dataset format representing the time spent by each user in each URL

User	URL1	URL2	URL3	URL4…
1	17	0	25	60
2	12	5	40	0
3	18	56	21	225
4	128	0	44	55
5	80	35	0	0
6 . . .	15 . . .	66 . . .	0 . . .	100 . . .

3.3 STUDENT DATA SET

3.3.1 Data Set Description

Any University or college management strives hard in fulfilling the aspirations of the students admitted. Amid the admitted students, some may be interested in jobs; some may be keen towards research and other towards higher education or entrepreneurship. In order to accomplish the goals, the authorities of the colleges need to understand the mindsets of the students and shape them accordingly. In order to understand the student‘s dreams, apart from the other methods such as questioning, the navigation behavior of the students also plays a vital role. In our proposed work the behavior and surfing characteristics of students are considered. This work elucidates the process of identifying the student behavioral pattern by tracing the web browsing behavior of the students hidden in the log files of GITAM university web server. The raw data is preprocessed, to generate organized information.

This data set contains the preprocessed and cleaned sessionized data for the main gitam.edu Web server (http://www.gitam.edu). The data is based on a random sample of students visiting this site for a 1 week period during September of 2013. An example of a line of data in a web log is*: #Date: 2013-09-09 12:39:57 192.168.23.31 - wsa1.gitam.edu GET http://b.scorecardresearch.com/p2? TCP_MEM_HIT/200 318*.This recorded information is in the format: date, time, host/ip user ”request url” statusbytes; it represents from left to right, the time of access (12:39:57 p.m. on Sept09, 2013 at a location 5 hours behind Greenwich Mean Time (GMT))the host IP address of the computer accessing the web page (192.168.23.31), the user identification number (-) (this dash stands for anonymous), request(GET http://b.scorecardresearch.com/p2), the unified reference locator (URL) of the web page being accessed (TCP), status of the request (which can be either 200 series for success, 300 series for redirect, 400 series for failures and 500 series for server error), the number of bytes of data being requested (318).

3.3.2 Preprocessing

Stage 1: Data cleaning

Algorithm : Data Cleaning
Input: Server_Log_Data
Output: Cleaned_HTTP_Log

Process:

```
    Step1: if data available in Server log then goto step 2
               elsegoto step 7
    Step2: Clean data by eliminating gap, image, audio and video files.
    Step3: Execute UserExtraction
    Step4:Goto step1
    Step5: Exit
```

Stage 2: Identify unique URL's and assign unique id to each url.

Algorithm : URLIdentification
Input: Cleaned_HTTP_Log
Output: URL_Log_Data
Process:

```
    Step1: Take each record from Cleaned_HTTP_Log file and extract ip address.
    Step2: Capture the string after GET
    Step3: Assign a unique URL_ID to the URL identified in step2 and Record each URL wise
               log data in a new line
    Step4: Repeat step 1 to step 3 until data is available
    Step5: Exit
```

Step 3: User identification

Algorithm : UserExtraction
Input: Cleaned_HTTP_Log
Output: User_Log_Data
Process:

```
    Step1: Take each record from Cleaned_HTTP_Log file and extract ip address.
    Step2: Convert ip address to domain name by reverse DNS lookup.
    Step3: Identify user by sending cookies
    Step4: Assign a unique USER_ID to the user identified in step3 and Record user wise log
               data in a new line
    Step5: Repeat step 1 to step 4 until data is available
    Step6: Exit
```

Step 4: Identifying the sequence of web pages traversed by each User

Algorithm : PageSequenceExtraction
Input: User_Log_Data, URL_Log_Data
Output: Page_Sequence_List
Process:

```
    Step1: Take each record from User_Log_Data and URL_Log_Data
    Step2: For each User_ID in User_Log_Data, Identify the time spent at each URL_ID in
               URL_Log_Data
    Step3: Record the time at each URL_ID and mark as '0' at unvisited URL_ID
    Step4: Repeat step 1 to step 3 until data is available
    Step5: Exit
```

Step 5: Converting the dataset into student versus access sequence

From the preprocessed data of algorithm 4, we extract the user and the sequence of pages he visited. The number of pages visited by each student varies from student to student. So the length of student access sequence also varies. By applying the concepts of merging and subsequence matching these variable length students access sequences are converted into equilength.

Step 6: Converting the dataset into student versus time spent in each page

From the preprocessed data of algorithm 4, we extract the time spent by the user in each page he visited. With this data we construct a user versus url vector matrix where each entry represents the time spent by the user in the corresponding url.

V. RESULTS AND DISCUSSION

Patterns extracted from the web logs help the analysts to grab the interests of the web users to enhance the design of the web page, the key element in the conduct of business and e-commerce. This paper presents a clear understanding about the preprocessing and pattern discovery algorithms applied on three different datasets. The analysis of web log involves the process of transfiguration and prediction of web log records to extract the hidden information and patterns. This process results in a great development of business intelligence.

VI. CONCLUSION AND FUTURE SCOPE

Patterns extracted from the web logs help the analysts to grab the interests of the web users to enhance the design of the web page, the key element in the conduct of business and e-commerce. This paper presents a clear understanding about the preprocessing and pattern discovery algorithms applied on three different datasets. The analysis of web log involves the process of transfiguration and prediction of web log records to extract the hidden information and patterns. This process results in a great development of business intelligence. The methodologies that are proposed in this paper can be further enhanced by considering a generalized technique that can be fit into various databases and ther by helping to derive useful patterns that are useful for mining big data.

REFERENCES

[1]. DoruTanasa and Brigitte Trousse, “Advanced Data Preprocessing for Intersites Web Usage Mining” , 1094-7167/04/$20.00 © 2004 IEEE

[2]. Xidong Wang, Yiming Ouyang, Xuegang Hu and Yan Zhang ,Discovery of User Frequent Access Patterns on Web Usage Mining, The 8th International Conference on Computer Supported Cooperative Work in Design Proceedings, sponsored by ieee 2003.

[3]. TahiraHasan, SudhirMudur and NematollaahShiri, A Session Generalization Technique for Improved Web Usage Mining, WIDM’09, November 2, 2009, Hong Kong, China.

[4]. Cooley, Robert, BamshadMobasher, and Jaideep Srivastava. "Data preparation for mining world wide web browsing patterns." Knowledge and information systems 1.1 (1999): 5-32.

[5]. Hussain, Tasawar, SohailAsghar, and Simon Fong. "A hierarchical cluster based preprocessing methodology for Web Usage Mining." Advanced Information Management and Service (IMS), 2010 6th International Conference on. IEEE, 2010.

[6]. Khanchana, P., and M. Punithavalli. "Web Usage Mining for Predicting Users’ Browsing Behaviors by using FPCM Clustering." International Journal of Engineering and Technology IACSIT 3.5 (2011).

[7]. Facca, Federico Michele, and Pier Luca Lanzi. "Mining interesting knowledge from weblogs: a survey." Data & Knowledge Engineering 53.3 (2005): 225-241.

[8]. Pallis, George, Lefteris Angelis, and Athena Vakali. "Validation and interpretation of Web users’ sessions clusters." Information processing & management 43.5 (2007): 1348-1367.

[9]. Sundari, M. Rekha, Prasad Reddy PVGD, and Y. Srinivas, "A Review on Pattern Discovery Techniques of Web Usage Mining". International Journal of Engineering Research and Applications. Vol. 4, Issue 9, September 2014.

[10]. Duarte Torres. Sergio, Ingmar Weber, and Djoerd Hiemstra. "Analysis of search and browsing behavior of young users on the web." ACM Transactions on the Web (TWEB) 8.2 (2014): 7.

AUTHORS BIOGRAPHY

Dr. Rekha Sundari.M was born in Amalapuram, East Godavari district in Andhra Pradesh. She completed her M.Tech from GITAM University, Visakhapatnam and Ph.D from JNT University Kakinada. She is presently working as Assistant Professor in Department of Computer science and Engineering, GITAM University, Visakhapatnam. Her research area includes Data Mining and Image Processing.

Srinivas Y is presently working as a Professor, in Department of Information Technology, Gitam University, Visakhapatnam. His research area includes Image Processing, Data Mining, and Software Engineering.

Prof. Prasad Reddy, P.V.G.D, was born in Rajahmundry, East Godavari district in Andhra Pradesh. He obtained his B.Tech, in Mechanical Engineering, from Andhra University and M.Tech, in Computer Science & Technology, and Ph.D, in Computer Engineering. He is presently the Head of Computer Science & Systems Engineering department, Andhra University, Visakhapatnam. His Research areas include Soft Computing, Software Architectures, knowledge Discovery from Databases, Image Processing, Number theory & Cryptosystems.