International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

An Novel approach on Pre-Processing Technique on Web log mining

Radha.M*1, K.Santhi*2

[1]Research scholar, computer science, Sri Ramakrishna College of arts & science for women, Tamil Nadu, India

[2]Associate professor, computer science, Sri Ramakrishna College of arts & science for women Tamil Nadu, India,[1,2]radhairjet2016@gmail.com

---

Abstract - *This work is inspired by the continued growth of Web-based information systems. This paper focuses on providing techniques for better data cleaning and feature extraction from the web log. After the process of data cleaning extract features from the accessible transactions in the log. In this paper seven different feature set was extracted from the transaction log entry. The experimental result shows that the feature extraction after data cleaning produces significant improvement in user navigation pattern analysis.*

Key Words: *Web Mining, feature extraction, cleaning, analysis, log file*

1. INTRODUCTION

Web-based organizations in their daily operations have reached astronomical proportions. Log data is usually noisy and ambiguous and preprocessing is an important process for efficient mining process. In the preprocessing, first the data cleaning process includes removal of records of graphics, videos and the format information, the records with the failed HTTP status code, robots files, and Session and user identification is performed. The next primary goal is to learn the user's navigation patterns and their use of web resources in web usage mining. Identifying the potential attributes and reducing the dimensionality of the data by not including irrelevant attributes are the major role of feature extraction. The assignment is to convert inconsistent length transactions into fixed-length feature vectors. The potential feature set extraction will lead to better understanding of the user navigation patterns in web server log files instead of taking into the consideration of whole instances in the log file.

2. RELATED WORK

Traditional student modeling techniques are inapplicable in these systems when tutors are overwhelmed by the huge volumes of sequential data generated as learners browse through the Web pages (Agrawal and Srikant, 1995) [1]. Web mining in education is not new. It has been applied to mine aggregate paths for learners engaged in a distance education environment (Ha, Bae and Park, 2000) [2]; e-articles for students based on key-word-driven text mining (Tang et al., 2000) [3], and to analyze learners' learning behaviors (Zaiane and Luo, 2001) [4].Yan Li [5] presented a detail algorithm method for web usage mining implementation of the data preprocessing system. After identifying the user session, the referrer based method is used to find the user's access path which is attached with an effective solution to the problems with proxy servers and local caching. Hussain .T, Sahail Asghar [6] proposed in the preprocessing level framework for web session cluster of usage mining. It covers the steps to prepare the log data and it converted into numerical data. Doru Tanasa [7] the research describes two main contributions to WUM process (i.e.) for preprocessing the web logs and a divisive with three approaches for the discovery of sequential patterns with a support. The algorithm used for the processing the web log records and obtaining the set of frequent access patterns have been implemented by huiping Peng[8]. An improved preprocessing expertise has been used by ling Zheng [9] for the purpose of solving some existing issues in traditional information preprocessing in web log mining. JIANG Chang-bin and Chen Li [10] says that, even if the statistical data are not enough and absence of visiting user history records, the web log data preprocessing algorithm based on collaborative filtering identifies the session flexibly and quickly.

3. PROPOSED FRAMEWORK

Once the data set was collected from the web server the raw log file has to undergo some preprocessing works. The log file has to be cleaned by removing the duplicate and missing value entries, the robotic file information also be included in the log file, images and the style sheet entries are also removed. Then the user identification based on the IP address is performed. The session identification based on the duration of entry and exit time is identified. Then the last stage in the preprocessing technique is the feature extraction process in which the feature sets are extracted using the methods

- First n and last n pages visited
- First n and last m pages visited
- Top n frequently visited pages
- Top n most time spends pages
- Top n infrequently visited pages
- Path traversal

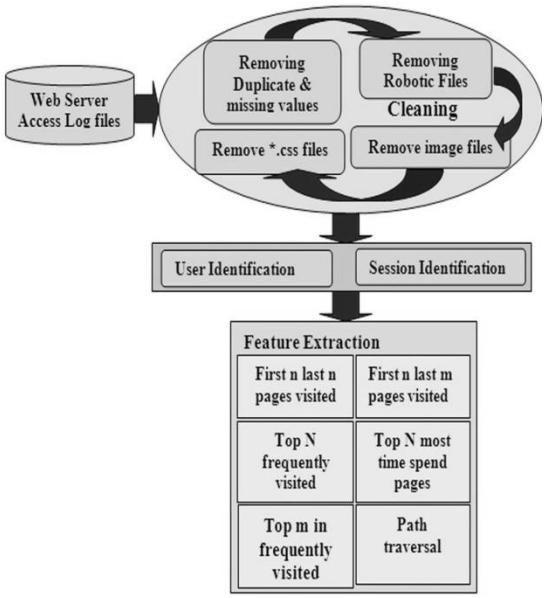

Fig-1: Overview of Proposed Framework

3.1 Framework of Data Preprocessing Process

The primary use of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consist of following steps

- Field extraction
- Data cleaning
- User identification
- Session identification

The data collected for usage mining is from diverse sources which represents the navigation patterns of various segments of the overall web traffic. Web server log does not exactly contain adequate information for inferring the behavior on the client side as they relate to the pages provided by the web server. Time Oriented Heuristic Algorithm for Session Identification which categorizes the users by session also improved from the traditional web based method. The experimental results clearly show the quality of the data, can be improvised by following the improved preprocessing techniques for analyzing, instead of using the traditional web based method. When a user submits a request to a web server that activity are recorded in the web log file. Log file ranges 1KB to 100MB. Example of a web log file

```
ip2283.unr - - [16/Nov/2005:00:01:03 -0500] "GET /dmcourse/dm.css HTTP/1.1" 200
155 "http://www.kdnuggets.com/ dmcourse/data_mining_course/ assignments/
assignment-3.html " "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
```

Fig-2: Example of Web Log File

3.1.1 Field Extraction

Each user entry is represented as a single line of the log file. The log entry contains many fields as discussed in the earlier section which has to be separated out for further processing. The filed extraction is the process of separating the field from the single line of the server log file. The server used different characters which work as separators. The most used separator character is ',' or 'space' character.

Delimiter based Field Extraction algorithm is given below.

Input : Log File

Output : DB

Open Log File

Read all fields contain in Server Log File

Separate out the Attribute using the delimiter

Extract all fields and Add into the Server Log Table (SLT)

Close

An example of Server Log Table (SLT) is shown in Table 1. Each and every section of the log entries has been separated to attributes for easy cleansing of unclean data form huge data source. The detailed description of the letters mentioned in the table is given below the table.

3.1.2 Data cleaning

This stage consists of removing the entire data track in Web logs that are ineffective for mining purposes as shown in figure 8. Graphic file requests, agent/spider crawling etc. could be simply removed by only seeming for an HTML file requests. Normalization of URL's is regularly required to make the requests consistent. In Data Cleaning, log entry involves irrelevant references to embedded objects like multimedia files which may not be necessary for analyzing purposes as given in table 1. Therefore such kind of useless entries has to be removed from log files before performing any analysis process. By performing Data cleansing process, errors and discrepancy will be discovered and removed to improve the quality of data. In Data cleansing process, following steps are performed

Algorithm for Data Cleaning

Read Entries in SLT

For each Entry in SLT

Read fields (Status code, method)

If Status code='200'and method= 'GET'

Then

Get IP_address and URL_link

If suffix.URL_Link= {*.gif,*.jpg,*.css} Then

Remove suffix.URL_link

Save IP_sddress and URL_Link

End if

Else

Next Entry

End if

3.1.3 User Identification

In this stage, the individual user is identified using their IP address. While reading the entry in the sever log table if the IP address is new then consider it as new user. If the IP address already exists but either the browser or operating system differs then it is also considered as different users. The algorithmic representation for the User Identification is given bellow.

Distinct User Identification algorithm using Server Log Table (DUI)

Read each entry in SLT

If an IP address not exist then

Consider the user as new user

End if

If IP address exists and the ((browser version or Operating System) is not exist) then

Consider the user entry as new user

Elseif

Next entry

3.1.4 Session Identification

The time duration spent on web pages are called Session. To identify the new session referrer-based method is used. When the IP address, browser version and operating system are same the referrer information should be taken. A new user session is identified if the URL in the Refer URI – field is a larger interval usually more than 30 minutes between the accessing times on this record.

Time Oriented Heuristic Algorithm for Session Identification (TOH)

For each entry in SLT

Sort the log data by IP address , agent and time Next entry

Read each entry in SLT

If the IP address and agent not exist then

If ($requested\_time_i$ - $requested\_time_{i-1}$) > Session Time Out or Session time-out does not belong to Session history then Increment the value of session by 1

Consider the user session as new

Endif

Endif

Next entry

After performing Data cleansing of the raw data set the quality of the cleaned data was validated with the existing clustering models namely Partitioned clustering, Hierarchical methods, Density based clustering and Sub Space Clustering. The results show on applying these models the quality of the datasets was highly improvised. So, the obtained cleaned datasets will be used for further personalization process.

3.2 Methodology for Proposed Feature Extraction

Feature extraction is the final step of the preprocessing stage which can be optional in many preprocessing of web data in the process. Feature extracted from the web data can be used effectively for further pattern analysis stage of the personalization

3.2.1 First n and Last n Pages

In this feature, set the first n and last n pages visited by the visitor in a transaction are taken into the consideration. This feature set is used to obtain more information in the data mining. Algorithm steps are given below

Step 1: Read the log file and Identify the navigation history of each visitor

Step 2: Count the no of first n and last n pages navigated by the visitor

Step 3: Sort the visited pages based on the count

Step 4: Determine highest first n and last n pages visited to be listed in the top row.

3.2.2 First n and Last m Pages

First n and last m pages visited in this attribute selection method each entry consist of first n and last m pages accessed are taken into the consideration. This type of grouping contains significant information to discriminate cadre of visitors. Algorithm steps are given below

Step 1: Read the log file and identify the navigation history of each visitor

Step 2: Count the no of first n and last m pages navigated by the visitor

Step 3: Sort the visited pages based on the count

Step 4: Determine highest first n and last n pages visited to be listed in the top row.

3.2.3 Top N-Most-Frequently Visited Pages

In this technique most frequently visited pages are identified. The Top N most frequently visited pages were chosen as attributes in this feature set. It was based on the notion that it would be feasible for classifying visitors browsing behaviors based on whether they visited a frequently visited page or not. Hence, this feature set was used. Algorithm steps are given below:

Step 1: Read the log file and count the number of unique visitors visited pages

Step 2: Sort the result based on the count

Step 3: Identify the top n most frequently visited page which has highest count

3.2.4 Infrequently Visited Pages

In this feature set the infrequently visited pages are selected for analyzing the reason why the visitors are avoid to view. Whether those infrequent page should be taken into consideration for a modification. Algorithm steps are given below

Step 1: Read the log file and identify the navigation history of each visitor

Step 2: the number of unique visitors visited pages
Step 3: Sort the visited pages based on the count

Step 4: Determine highest first n and last n pages visited to be listed in the top row.

3.2.5 Top N-Most-Time Spend Pages

This feature set has the same attributes as those in Top N-Most-Frequent visited pages. In this feature set, an attribute value is the amount of time spent in seconds by the visitor on that particular frequently visited page. The duration was calculated by taking the time difference between two consecutive page requests. This feature set is based on the conjecture that time spent on frequently visited pages might be a very distinguishing factor to categorize visitors. Algorithm steps are given below

Step 1: Read the log file and Count the number of unique visitors visited pages

Step 2: Using the visitor entered time and the exiting time of a particular page

Step 3: Calculate the total time spend on the particular page

Step 4: List the pages in the order of the most time spend pages

Step 5: Identify the top n highest pages

3.2.6 Path Navigation

This approach of attribute selection method is used to analyze how the users visited a particular web page. The association existed among the web page navigation is considered in analyzing the user browsing patterns. Algorithm steps are given below

Step 1: Read the log file and select the target page to be analyzed

Step 2: Cluster the visitors based on the last page they visited

Step 3: Identify the most frequent path chosen for navigating the concern page

3.2.7 First and Last Pages Visited

The pages which are viewed first and last are identified. Whether the visitors enter on the website by the main page or they tried from a different page. Inferring, why visitors left following reading these pages can be difficult it could be that these pages comfortable their information requirements, or that they became irritated by the time they arrived at these pages. Algorithm steps are given below

Step 1: Read the log file; Count the no of unique visitors visited first and last page

Step 2: List them based on the highest count ratio

Step 3: Categorize the most visited first and last pages.

4.EXPERIMENTAL RESULTS

4.1 Performance of Data Cleaning Process

The data set cleaning process is performed with the traditional preprocessing based on website topology with the proposed DUI and TOH model.

Table-2: Web log file result for Web Based and proposed DUI & TOH algorithm

Web log file information	Web based Model	DUI & TOH Algorithm
Record in original web log file	39974	39974
Records in cleaned log file	14678	20065
No. of. Users Identified	1508	2307
No. of. Session Identified	678	1787
Reduction (%)	36.7	50.2
Accuracy (%)	86.8	92.7

The table 2 shows original data size is 39974 Instances after applying the preprocessing process of proposed model such as removing duplicates, robotic files and image files it was reduced to the percentage of 50.2. The number of users identified was 2307, the number of sessions identified was 1787 and the Accuracy 92.7. When we consider the same data set in traditional based web topology models, the record cleaned log files are 14678, the percentage of reduction is 36.7, number of users identified is 1508, number of sessions identified is 678 and the Accuracy is 86.8 which is much less compared to the proposed model as shown in figure 10.

Table-3: Clustering Model results for Web Based and DUI&TOH algorithm accuracy

Clustering Models	Web based Model Accuracy	DUI & TOH Algorithm Accuracy
Farthest First	75.67%	85.09%
Make Density Based	86.89%	92.73%
COBWEB	71.06%	83.56%
K – Means	74.87%	86.32%
EM algorithm	82.05%	90.78%

Table-4: Comparison for Web based and DUI & TOH algorithm in User and Session Identification

Web log file information	Web based Model	DUI & TOH Algorithm
No. of. Users Identified	1508	2307
No. of. Session Identified	678	1787

4.2 Performance Comparison of Proposed Feature Extraction

This proposed method uses two different log server files for feature extraction. They are online shop log file and the kdlog file. The experimental result was conducted with the help of the web log server open software tool. The result shows the top 10 most frequently visited pages, top 10 most frequently time spend site, the frequent paths the visitors used. The way how the entered the pages and Top 10 site areas are visited

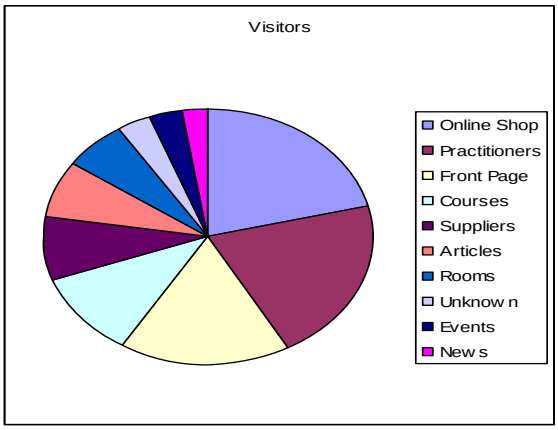

Fig-2: 5 Top 10 frequently visited paths

The Table 5 and fig 2 depicts the top 10 frequently visited paths by the visitors from the table it is observed that the path visited to online shop is higher than the remaining pages. Next top frequently visited paths are the practitioners and the front pages. The news page comes under the less interest by the visitors. This information can be used to observe the most used paths by the visitors. This feature set extraction was thought that it would be feasible to classify browsing behaviors based on whether they visited a frequently visited page or not.

Table-5: Top 10 frequently visited paths by the visitors

Top 10 frequently visited paths and no of vistiors	
Visit Path	**Visitors**
Online Shop	498
Practitioners	487
Front Page	390
Courses	250
Suppliers	188
Articles	171
Rooms	141
Unknown	80
Events	77
News	59

Table-6: Summary of the page entry

Summary of page entry		
Direct Requests	35%	1,039
From Linking Sites	10%	286
Direct Front Page	14%	410
From Search Engines	55%	1,629
Front Page	19%	551

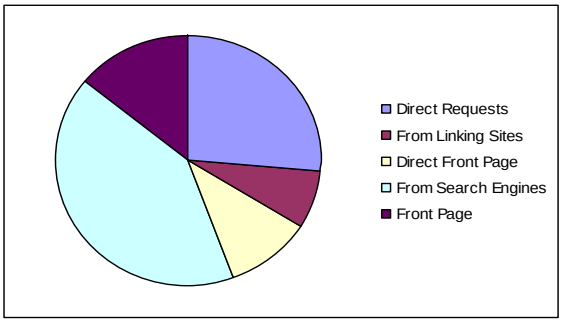

Fig-5: Visitors entered the home page

The Table 6 and fig 5 shows the way how the visitors entered the home page of the website. The search engines

Top 10 site areas visited		
Page	**Site Area**	**Visitors**
Front Page	Front Page	590
View Modality (Practitioner)	Practitioners	193
Rooms/Space for Rent Directory	Rooms	172
Supplier Directory	Suppliers	143
Practitioner Directory	Practitioners	123
Product Directory	Online Shop	120
View Product: item	Online Shop	110
Quick Find	Unknown	107
Links Page	Unknown	104
Course Directory	Courses	96

played a major role in page entry than the direct request and the front page request. In this table the request through which the visitors entered the page is listed. The highest page entry request comes from the search engines.

Table-7: Top 10 site areas visited

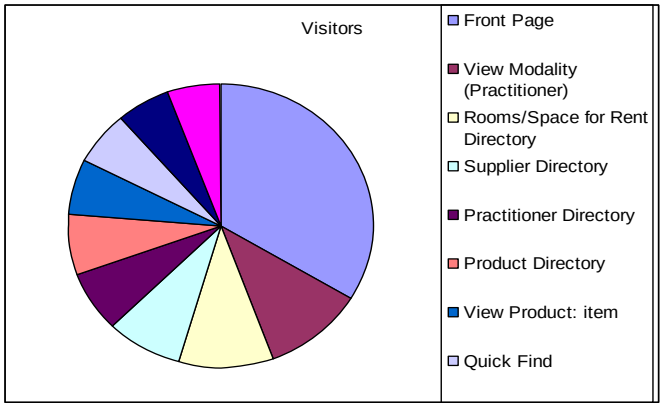

Fig-6: Top 10 frequently visited areas

In the table 7 and fig6 shows 10 most frequently visited site areas were selected as attributes in this feature set. The front page is the most visited site area by the visitors. The Course directory is least used. Online shop is the site area used to visit product directory and the product item view. The site areas mostly visited are used to identify what type of visitor used it and the overall outlook of the table shows the most of the visitors choice is front page. A site area is a logical grouping of urls within your site, this grouping is then given a name. For instance you may have a Products area of your site, it is useful to know how many visitors went to that area of your site.

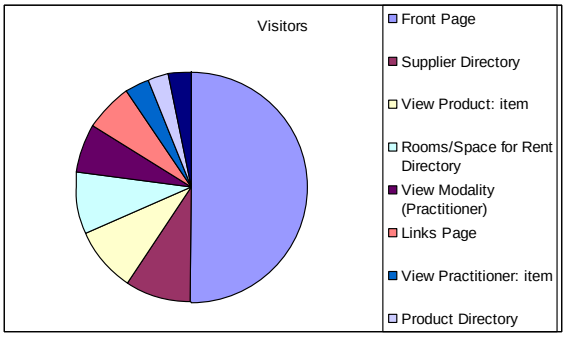

Fig-7: Top 9 entry pages by the visitors

From the table 8 and fig 7 its observed that the root page is the site area the most visitors used for entering the online shopping. The supplier's uses there own link page. The online shop is used to view the product details and the practitioners alone used the view modality page directly.

In the table 9 the most frequent time spend page is listed and it is observed that the visitors spend more time on practitioners page than the other pages. Next online shop was viewed for a long time by the visitors. The articles, news and events are taken less interest among the visitors. It is necessary to consider and determine why the page is not much attracted by the visitors. In this feature set, an attribute value is the amount of time the visitor spent on that particular frequently visited page, i.e. the attribute value is the time spent in seconds. The duration can be calculated by taking the time difference between two consecutive page requests. The attribute value \0" indicates that this particular page was not visited by the visitor in that transaction

Table-8: Top 9 entry pages by the visitors

Top Entry Pages by the visitors		
Page	**Site Area**	**Visitors**
Front Page	Front Page	551
Supplier Directory	Suppliers	103
View Product: item	Online Shop	101
Rooms/Space for Rent Directory	Rooms	94
View Modality (Practitioner)	Practitioners	75
Links Page	Unknown	74
View Practitioner: item	Practitioners	38
Product Directory	Online Shop	34
Events Directory	Events	33

Table-9: Top 10 Most Frequent Time spend pages

Top 10 Most Frequent Time spend pages				
Name	**Unique Visitors**	**Time Spent In Area**	**Visitors Entering**	**Visitors Exiting**
Practitioners	729	1d 14m 47s	579	609
Online Shop	690	18h 9m 11s	575	594
Front Page	590	7h 19m 7s	551	423
Courses	421	14h 45m	333	321
Unknown	339	3h 54m 10s	113	141
Suppliers	275	9h 37m 23s	234	223
Rooms	253	11h 1m 43s	169	200
Articles	223	3h 35m 1s	185	188
News	128	2h 12m 19s	79	73
Events	124	3h 21m 14s	92	95

The table 10 shows the feature set extraction of online shop path the visitors mostly performs direct access to online shop path. The remaining paths used for navigation are front pages, suppliers, practitioners and news.

Table-10 : online shop path

Path selected to view online shop	
Visit Path	**Visitors**
Online Shop	498
Online Shop->Unknown->Online Shop	10
Front Page->Online Shop	7
Suppliers->Online Shop	6
Front Page->Online Shop->Front Page	6
Practitioners->Online Shop->Practitioners	5
Online Shop->Practitioners	5
Online Shop->News->Online Shop	4
Online Shop->Contact Us->Online Shop	4
News->Online Shop	4

The table 11 depicts the various paths navigated by the visitors to access the practitioner's page. The direct page access to the practitioners scores high than the other navigation paths. Next course page acts as the

Table-11: Visitor path selection for practitioners

Path Listing		
Pages/Files	**Visitors**	**Percentage of visitors**
/	666	16.20%
/software/	187	4.55%
/jobs/	136	3.31%
/datasets/	64	1.56%
/companies/consulting.html	69	1.68%
/dmcourse/data_mining_course/	53	1.29%
/software/suites.html	64	1.56%
/software/visualization.html	52	1.26%
/news/2005/n21/	57	1.39%
/software/text.html	47	1.14%

Table-12: Visitor path selection for practitioners

Visitor path selection for practitioners	
Visit Path	**Visitors**
Practitioners	487
Courses->Practitioners	26
Courses->Practitioners->Courses	10
Practitioners->Unknown->Practitioners	8
Rooms->Practitioners	8
Practitioners->Rooms	7
Practitioners->Courses	6
Practitioners->Online Shop->Practitioners	5
Online Shop->Practitioners	5
Practitioners->Unknown->Practitioners->Unknown->Practitioners	5

In the table 12 shows the path listing of the visitors in Kdlog file. It is observed that root page contains the highest number of visitors, and then the software details and the jobs are visited often. Through the software page the visitors accessed the suites, visualization and text pages of the website. This information combined with the visitor location will help to determine their browsing pattern. This technique helps in doing a content-based feature extraction entry page for accessing the practitioner's path. This information will make it easier to differentiate between the types of visitors and their interests.

5. CONCLUSION

This paper summarizes on both data cleaning and feature extraction they come under data preprocessing web log file format, its type and location. Log files usually contain noisy and ambiguous data. In Data cleaning it involves removal of unnecessary data from log files. The cleaned log file entries are processed by the clustering techniques to perform pattern analysis and to identify the quality of data. The feature extraction technique plays a vital role in the web log file access.

The main aim of this work is to identify the feature sets which are useful for categorizing the visitors based on their navigation behavior. The query of whether infrequently visited pages carry useful

discriminatory information is left for further work. The most frequent time spend pages feature set is based on the inference that time spent on frequently visited pages may be a very unique factor to categorize visitors. Extracting this kind of feature set plays a significant role with extension to pattern analysis in user navigation pattern.

6. REFERENCES

[1] Agrawal, R., and Srikant, R. (1995). Mining Sequential Patterns. In Proc. of the Eleventh International Conference on Data Engineering (ICDE), Taiwan. Pp 3-14.

[2] Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan-Kaufmann Academic Press, San Francisco.

[3] Tang, C.; Lau, R.W.H.; Li, Q.; Yin, H.; Li, T.; and Kilis, D.(2000). Personalized Courseware Construction Based on Web Data Mining. In Proc. of the First International Conference on Web Information Systems Engineering (WISE 2000) vol.2, Pp. 204-211.

[4] Zaiane, O. and Luo, J. (2001). Towards Evaluating Learners' Behavior in a Web-based Distance Learning Environment. In Proc. of IEEE International Conference on Advanced Learning Technologies, Pp 357-360, Madison, WI.

[5] Tang, Y. T. and McCalla, G. (2001). Student modeling for a Web based Learning Environment: a Data Mining Approach. Department of Computer Science, University of Saskatchewan, Canada.

[6] Yan Li, Boqin Feng, Qinjiao Mao, "Research on Path Completion Technique inWeb Usage Mining", International Symposium on Computer Science and Computational Technology, 2008

[7] Hussain, T., S. Asghar, et al. 2010. Web Usage Mining: A Survey on reprocessing of Web Log File. IEEE, International Conference on (ICIET),pp. 1- 6

[8] Huiping Peng. 2010. Discovery of Interesting Association Rules Based On Web Usage Mining. IEEE conference, pp. 272-275.

[9] Ling Zheng, Hui GUI and Feng Li. 2010. Optimized Data Preprocessing Technology For Web Log Mining. IEEE International Conference On Computer Design and Applications (ICCDA), pp. 19-21.

[10] JING Chang-bin and Chen Li. 2010. Web Log Data Preprocessing Based On Collaborative Filtering. IEEE 2nd International Workshop on Education Technology and Computer Science, pp. 11