Web Usage Analysis of University Students to Improve the Quality of Internet Service

ANANDAN BELLIE

***Abstract*—Internet facility is one of the important infrastructure requirements to improve the quality of education. However, the quality of the internet service in terms of availability and accessing speed usually get suffered due to the unnecessary web contents access during work hours. There is a huge amount of data generated from the internet user activity in the web browser history is of no use without finding interesting pattern out of it. Data mining is the emerging research area to extract interesting patterns from huge amounts of Web data. There is an enormous amount of web log data generated every day in a personal computer history. As the number of users accessing internet is increasing day by day the issue of internet quality has to be addressed. This issue can be solved only by knowing the existing users behavior pattern. The Data mining based analysis is the best alternative to overcome secrecy issues of manual method monitoring. In this paper, the randomly chosen laboratory computers web browser history data collected to make clustering analysis model in WEKA, the free data mining tool. The data mining clustering technique has been chosen to group the similar characteristics students because of the enormous amount of web history log data. This model can be helpful to the system administrator to monitor and control the internet access. This analysis result may help the university management to plan for the future internet infrastructure requirements. As the data is collected from the individual user's web history, this analysis can also be used to customize the web page personalization purpose.**

Keywords--Web data mining, Browser history, Clustering, Simple KMeans, and Quality of service.

I. INTRODUCTION

Students are accessing the internet service for different purposes for their day to day learning activities. There are many services of internet like FTP, HTTP and mail are there for the students to use. But most of the time, the students are using the http service for the activities like browsing and downloading. As the internet usage is increasing day by day, the effect of usage need to be analyzed for the improvement of the service. This can be realized only by knowing the existing behavior patterns of the users. These patterns are like students accessed site name, time, frequency of accessing and the duration they stayed in that sites are useful for the later enhancement of the internet service quality. This analysis can also be used to know the level of understanding about the real use of the internet among the students community. Their way of accessing the Internet either from the referral site or

direct visit by entering the name in the web browser address bar. Another benefit of this analysis can help the department to know the students interest in using the internet service for their teaching and learning activities. The computer lab technical staff can now have a better control over the lab internet services by blocking few of the unwanted websites in work hours. The Administrator can monitor and control the bandwidth requirements. The management may plan for the further infrastructure requirement for the internet service.

The reason for this research is to reduce difficulties in analysis the huge amount easily available historical data from the web browser to improve the internet service quality. Keeping the secrecy issues and the easy accessible of data in mind, the users access web browser log is taken from the commonly used computers which is assigned toy many different levels of students of the department lab. Internet access facility is one of the important requirements in recent days for a day to day teaching and learning activity. World Wide Web has become a necessary platform to upload and download educational information but due to huge, diverse, and dynamic data, web data research is one of the useful areas in recent days. As a result, web users are of different groups are exists for different needs. However the information which drawn from the net is not much relevant to the user requirements and this leads to the problem of information mess which will degrade the internet service quality. As the user interacts with the web of various sites for various links, there is a wide diversity of user's access behavioral patterns exit, which can be easily analyzed by the automatic analysis tools to improve the Internet service quality This Analysis result, can also be used by the administrator to increase the download and upload speed and to know the existing user's behavior. Data mining is one of the multidisciplinary filed is used in this paper for the web data usage analysis. There are two types of data mining types, first type is predictive and another one is descriptive. The descriptive type method category clustering algorithm is used in this paper. Since the data from the history log file is huge, the clustering technique is the most appropriate one .The Simple KMeans Algorithm is used to group the similar characteristics users. There are 100's of records generated every week from an individual users computer web browser log is taken for analysis. The standalone personal computer history data of randomly chosen users log file of around 100 records each from 10 different computers from two computer lab is taken for analysis in WEKA tool.

Manuscript received June 2015.
*Anandan Belli**e**, Department of Software Engineering, College of Computing and Informatics, Haramaya University Haramaya , ,Ethiopia Ph.0251-932407953*

The rest of this paper is organized as follows. Section 2 is about the literature survey. Section 3 is about the methodology applied. Section 4 is the result analysis and section five for conclusion and scope for future study.

II. RELATED WORK

There are many papers published about the web server log file analysis but only few papers may be published on web browser log file data clustering. These data are easily accessible web browser history from the commonly used many students computers. Three types of mining can be done with the web data.

Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables [4]. This paper deals with a study of different techniques and pattern of content mining and the areas which has been influenced by content mining.[8]

Web Structure Mining: Web structure mining is based on the link structures with or without the description of links. The goal of web structure mining is to generate structured summary about websites and web pages.[9]

Web Usage Mining: This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web page. Web server gathers this information automatically into the Access Log file [10].Web usage mining is used to discover interesting user navigation patterns and can be applied to many real-world problems, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behavior studies, etc[4].

This paper [1] shows how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior. In this paper the PD-Tree algorithm was used in frequent tree mining task. Paper [6] suggest web mining may be organized into the following subtasks:

• Resource discovery. Locating unfamiliar documents and services on the Web.
• Information extraction. Automatically extracting specific information from newly discovered Web Resources
• Generalization. Uncovering general patterns at individual Web sites and across multiple sites

Raymond Kosala and Hendrik Blockeel [7] suggest decomposing Web mining into these subtasks, namely:

1.Resource finding: the task of retrieving intended Web documents.
2. Information selection and pre-processing: automatically selecting and pre-processing specific information from retrieved Web resources.
3. Generalization: automatically discovers general pat- terns at individual Web sites as well as across multiple sites.
4. Analysis: validation and/or interpretation of the mined Patterns

Supinder Singh, Sukhpreet Kaur [5] surveyed different functionalities of web Usage Mining which are related with web files.As every data mining task, the process of Web usage mining also consists of three main steps:(i) preprocessing, (ii) pattern discovery and (iii) pattern analysis. [1].

Research directions of [3] describes that the set of frequent patterns derived by most of the current pattern mining methods is too huge for effective usage. There are proposals on reduction of such a huge set, including closed patterns, maximal patterns, approximate patterns, condensed pattern bases, representative patterns, clustered patterns, and discriminative frequent patterns, were introduced.

Cluster analysis in high-dimensional space is a challenging problem. Since it is easy to compute frequent patterns in subsets of high dimensions, it provides a promising direction for high-dimensional and subspace clustering [2].Rajni pamnani and Pramila chawan provided a survey and analysis of web usage mining systems and technologies. They also discussed about an application of an online recommender system that dynamically generates links to pages that have not yet been visited by a user[11]

III. RESEARCH METHODOLOGY

Data collection

Due to the secrecy of the internet users access information accessing the server data is not easy, the simple and easy way to get the web usage data is from the Personal Computer history of data are collected. The student's lab computers used by different study years of students data is taken here for analysis. 100's of history record created every day and thousands of records created every week in the history file taken for analysis.
There are two major internet browsers data collected. The chrome browser data is taken for analysis because of the easy access of the time information which is present in the history file; snapshots of this data are shown in the figure 1.

Analysis steps

The following figure 1 shows that the randomly chosen computers from multiple users of the individual computers Browser is preprocessed and input to the WEKA tool for analysis. In this paper cluster analysis means applying the Simple KMeans algorithm to group similar characteristics users from the web browser history. Before applying this data to the mining process the manual and WEKA tool preprocessing phase ah been done so that, the quality result can be achieved

Tools used

WEKA is a open source data mining toll used for analysis purpose. And the result of this also compare with the Excel spreadsheet software package.
Figure 1 shows the process of Web log clustering process of web usage mining.

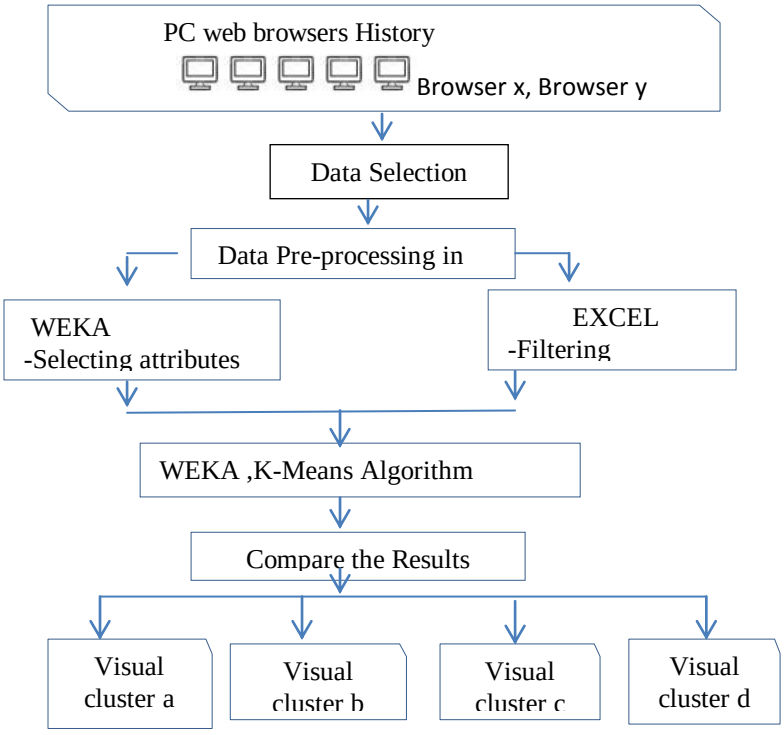

Fig 1. Analysis procedure

Data are collected from Google Chrome and Mozilla Firefox browsers history of randomly chosen computer from two different laboratories. Only Google Chrome history data are selected and used to the cleaning phase. There are two tools used for cleaning, one is by applying the excel features and then the WEKA tool. After a cleaning process the clustering algorithm applied to form cluster formation. It is depicted in Figure 1.

58. 12:15 PM
https://www.google.com.et/webhp?sourceid=chrome-instant&r
www.google.com.et
59. 12:14 PM
Sign in - Google Accounts
accounts.google.com
60. 12:12 PM
Welcome to Facebook - Log In, Sign Up or Learn More
www.facebook.com
61. 12:08 PM
YouTube
www.youtube.com
62. 12:08 PM
عيون العرب
www.alarabeyes.com
64. 11:51 AM
https://www.facebook.com/photo.php?fbid=753648714733305&set=gm.883297705063780&type=1&theater
www.facebook.com
66. 3:30 AM
Welcome to Facebook - Log In, Sign Up or Learn More
ww.facebook.com

Fig 2. Sample data from the browser history log.

Using EXCEL add-in Kutools the row wise data is transformed to column wise data and the feature selection of the data are as show below.

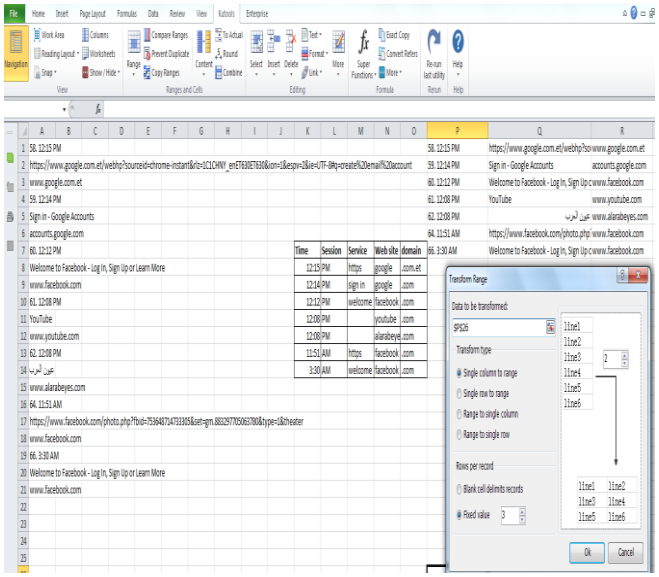

Fig 3. Pre-processed data in Excel

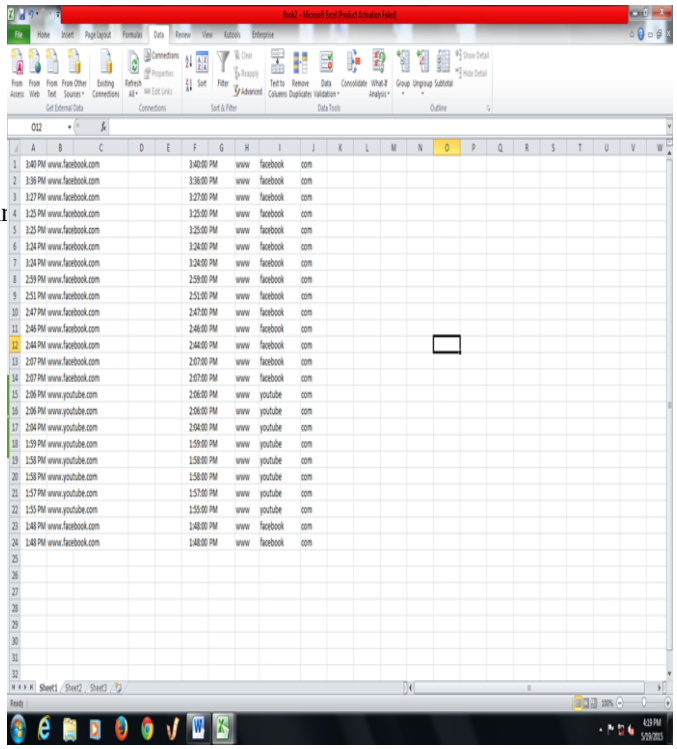

Fig 4. Selected sample data, after Text to Columns data conversion view.

IV. RESULT ANALYSIS

There are 880 records from two computer laboratory history records are taken for analysis and their results are shown below.

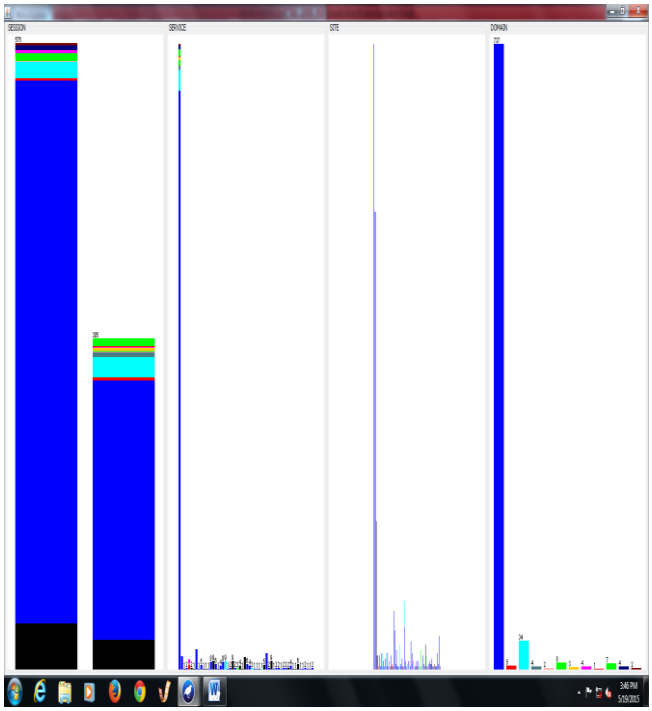

Fig 5. Four attributes distribution view

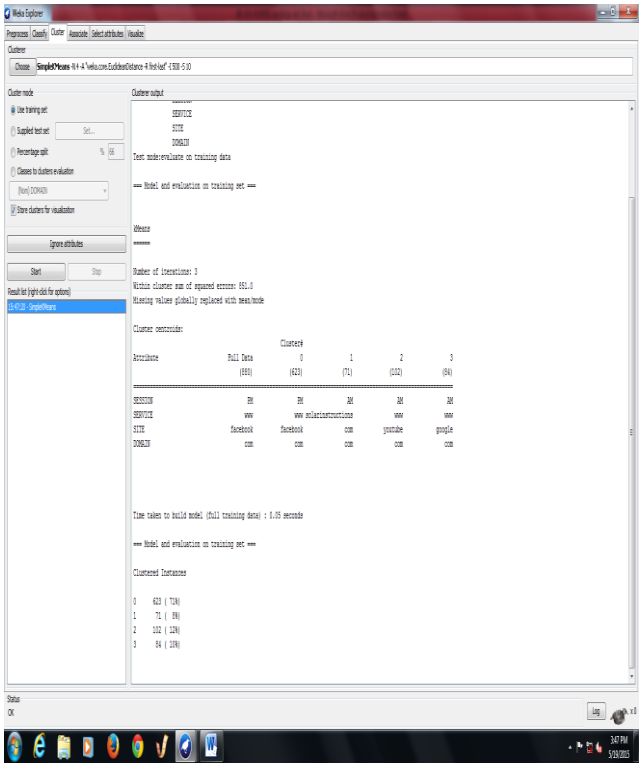

Fig 6. Simple k-means cluster output

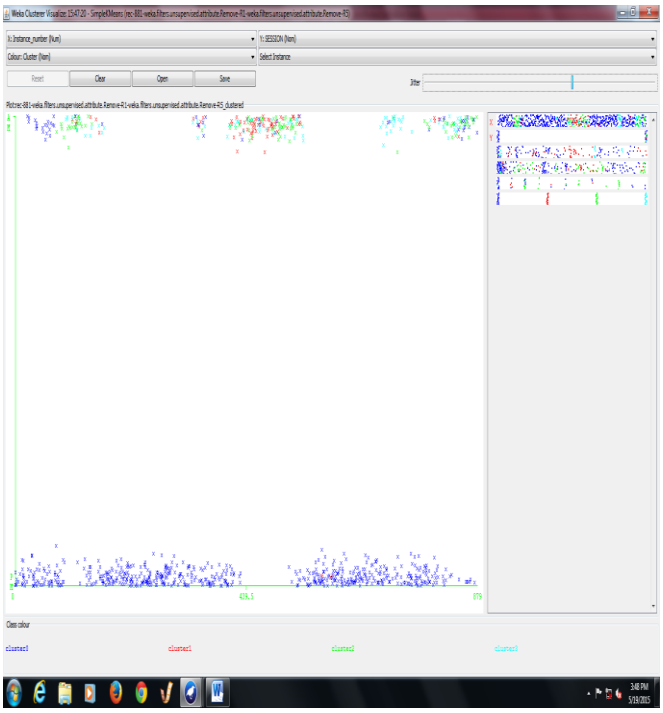

Fig 7. Visual Clusters view

Randomly selected personal computers among many computers from two departments' laboratory web browser history of 880 records are selected and input to the data analysis task. This selected attributes data is preprocessed further in excel and WEKA before mining process. Assumed 4 clusters distributions with settings of seed value 10 and the number of iterations 500 was show in the above Fig 7.

TABLE I. SUMMARY OF THE STUDENTS WEB SITE NAME ACCESS

Site Name	Usage.No of Records	In %	Rank
Facebook	623	70.79	1
Youtube	102	11.59	2
Google	84	9.54	3
Other sites	71	8.06	4

TABLE II. SUMMARY OF THE WEB USAGE

Category	First Preference
Session	AM
Service	HTTP
Site	Facebook
Domain	com

In the above summary information of the web usage web site name access Table I, the highest percentage is Facebook. In Table II the usage in AM session is high compare to PM and HTTP service more than the other service like FTP and Mail . Facebook web site name is more than other sites and .com domain access is high compare to others like educational and network domain is also shown in Table -II

Web browser log file is like background recoding of

the users activity. Mozilla Web browser of many computers of different students data are collected for the manual preprocessing and it is used for clustering. There are many websites names are not fully available and the missing complete information of each activities are removed using excel filtering tool.One week history data are collected for analysis and the following parts of the data are considered for analysis purpose.

- Web service type like FTP or HTTP data contents like the text, images, audio or video or information retrieved from web history data are collected.
- Time slot of AM or PM data represent the time they accessed the content is organized
- Usage data represent a Web site's name such as a Facebook, Youtube
- Domain like .com .edu

As the web server log is generally of huge in size in a case of large network and due to secrecy and sensitivity issues, few of the network computers personal web browser data are taken in this paper for analysis purpose.After the preprocessing steps the data taken from the Excel is stored in csv file format for WEKA tool. There is also one more preprocessing steps in WEKA tool to select the required attributes for clustering. Then the file is loaded in to WEKA tool for the K-means clustering algorithm. The result obtained from the WEKA tools are, Four types of group are identified from four attributes Time session as AM or PM and internet service like HTTP,FTP or E-Mail ,the website name and the domain like commercial, educational.

The results of the analysis show that 70% of the students are accessing the Facebook and other information shown above. The system administrator can block the site and other information considered to be unwanted access during work hours and may cache the frequently accessed useful pages for future use.

V. CONCLUSION AND FUTURE STUDY

There are 880 sample students web activity of browser log data results shows that majority of the students are accessing the Facebook and Youtube sites followed by Google in most of the time. It shows the students interest in social network site for various activities. This analysis report can be used to counsel and encourage the students to access the educational materials through this web site to avoid diversion of accessing unwanted information. If the social network based sites are well designed to access the educational materials then use this social network to motivate the students. The Youtube users are the second highest of 11 % compare to other sites. Rarely students accessed educational domain directly from the internet. This results helps the administrator to divert the band width requirements to other user location of the university if the users are not really using the internet for educational related activities during study time.

REFERENCES

[1] Renáta Iváncsy, István Vajk Frequent Pattern Mining in Web Log DataActa Polytechnica Hungarica Vol.3, No. 1, 2006.
[2] Jiawei Han Hong Cheng Dong Xin Xifeng Yan,Data Min Knowl Disc(2007)15:55–86 DOI 10.1007/s10618- 006-0059-1 Frequent pattern mining: current status and future directions.
[3] F. Bonchi, F. Giannotti, C. Gozzi, G. Manco, M. NanniD. Pedreschi, C. Renso, S. Ruggieri Web log data warehousing and mining for intelligent web caching, Data & Knowledge Engineering 39 (2001) 165-189
[4] A.Jebaraj RatnaKumar, An Implementation of web personalization using web mining techniques, Journal of Theoretical and Applied Information Technology
[5] Supinder Singh, Sukhpreet Kaur Study on Various Web Mining Functionalities using Web Log Files
[6] Oren Etzioni The World-Wide Web: Quagmire or Gold Mine? 66 November 1996/Vol. 39, No. 11 Communication of the ACM
[7] Raymond Kosala and Hendrik Blockeel arXiv:cs/0011033v1 [cs.LG] 22 Nov 2000 Web Mining Research: A Survey, SIGKDD Explorations.2000 ACM SIGKDD, July 2000. Volume 2, Issue 1- page 1
[8] Faustina Johnson , Santosh Kumar Gupta Web Content Mining Techniques: A Survey
[9] R.Malarvizhi, K.Saraswathi Web Content Mining Techniques Tools & Algorithms A Comprehensive Study, International Journal of Computer Trends and Technology (IJCTT) – volume4 Issue8–August2013
[10] Mehak, Mukesh Kumar, Naveen Aggarwal, Web Usage Mining: An Analysis , Journal of Emerging Technologies in web intelligence , Vol 5, No. 3, August 2013
[11] Rajni Pamnani and Pramila Chawan, " Web Usage Mining: A Research Area in Web Mining"

ACKNOWLEDGMENT

By GOD's Grace I dedicated this work to my parents, my family and all.

Anandan Belllie holds a Master in Software Engineering from Anna university of Technology affiliated college , Coimbatore. From year 2012, he has been working as a Lecturer in Haramaya University at the Department of Software Engineering. He has more than a decade of teaching experience.. He worked as Assistant Professor in various affiliated engineering colleges of Anna University. He alone published two international journal papers on data mining. His research thematic areas are in data mining and software engineering.

