Web Usage Characterization for System Performance Improvement

Alehegn Kindie[1(✉)], Adane Mamuye[1], and Biniyam Tilahun[2]

[1] Faculty of Informatics, University of Gondar, Gondar, Ethiopia
alehegn12@gmail.com
[2] Department of Health Informatics, University of Gondar, Gondar, Ethiopia
http://www.uog.edu.et

Abstract. Web usage mining discovers patterns of user behaviors from web log files. In this study web usage mining is employed to identify business-critical and non-business critical web traffics in University of Gondar. Apriori and FP tree algorithms are applied to extract the web browsing behavior in terms of frequently accessed sites along with their web traffics. Our research findings can be used as an input for bandwidth management and system performance improvement.

Keywords: Web usage characterization · Web usage mining
Pattern discovery

1 Introduction

World Wide Web is a global village and rich source of information [1]. Million of users engage with different web services and their browsing behavior evolve as rapidly as the business landscape underlying it. This evolution, however, is difficult to observe because of the users behavior. As a result, accurate picture of how users engage with the web is lacked [11]. Web usage mining, hence, aims to discover interesting and frequent user access patterns and trends from web [5,7]. The identified usage patterns are used for web usage characterization, bandwidth management, system performance improvement and web personalization [2–4].

Several approaches have been forwarded to extract knowledge from the Web. There are three broad categories of web mining: web content mining, web structure mining and web usage mining [2–4]. The web log files are major source of data for web usage mining. The users find relevant information easily and want to download the web resources with least amount of time [2,7]. System administrators want to create new knowledge from the usage pattern and want to improve the performance of the system. It is not possible to manage users access and improve the system performance without knowing their behavior. Web usage mining provides the key to understand interest of users and web traffic behavior which can in turn be used for network transmission and load balancing policy development [12]. For any organization, one can categorize web users

F. Mekuria et al. (Eds.): ICT4DA 2017, LNICST 244, pp. 238–245, 2018.
https://doi.org/10.1007/978-3-319-95153-9_23

as business-critical web user and non-business-critical web user or bandwidth abuser. For network and system performance improvement, adding more bandwidth can be one of the solution. However, since there are bandwidth abusers, adding bandwidth could not be a solution. Allocating required bandwidth to the business-critical network traffic and reducing non-business related network traffic (unwanted traffic) are the primary solutions for network and system performance improvement. This can be done using load balancing and task priority techniques [12]. Therefore, with the right policies in place, there is a possibility to improve the system performance.

In this paper, we investigated the web usage behavior of University of Gondar (UoG) community. The remainder of the paper is organized as follows. Section 2 illustrates the methodology followed in our study. Section 3 discusses the results achieved. While conclusions is given in Sect. 4.

2 Methodology

2.1 Data Collection

A total of 90,058,655 unprocessed web log data was collected from UoG proxy server.

2.2 Web Usage Mining Process

Our web usage mining process is divided into three major activities, namely data preprocessing, pattern discovery and pattern analysis, as presented in Fig. 1.

Data Preprocessing: The unprocessed web log data is cleaned using Glogg and Data Preparator. Then the data is clustered into students and staffs based on their VLANs. A total of 170,821 clean web log was used for experimentation purpose.

Pattern Discovery: Statistical analysis and association rule mining techniques are used for pattern discovery. Statistical Analysis is employed to identify access frequence [8]. Association rule mining algorithms, namely Apriori and Frequent Pattern (FP) tree, are used to extract interesting patterns. Though there are a number of techniques, as noted by Han and Kamber [10], the most widely used algorithms for association rule discovery are Apriori and FP tree. Since Apriori algorithm repeatedly scans all web log data, it takes long running time. While FP tree algorithm is faster than Apriori algorithm it scans the web log data only twice. On the other hand, FP tree algorithm takes only binary value. Even if the attribute VLAN has more than two binary values; it is not considered in FP tree algorithm experiment. As a result, the Apriori algorithm discovers an interesting pattern from each VLAN.

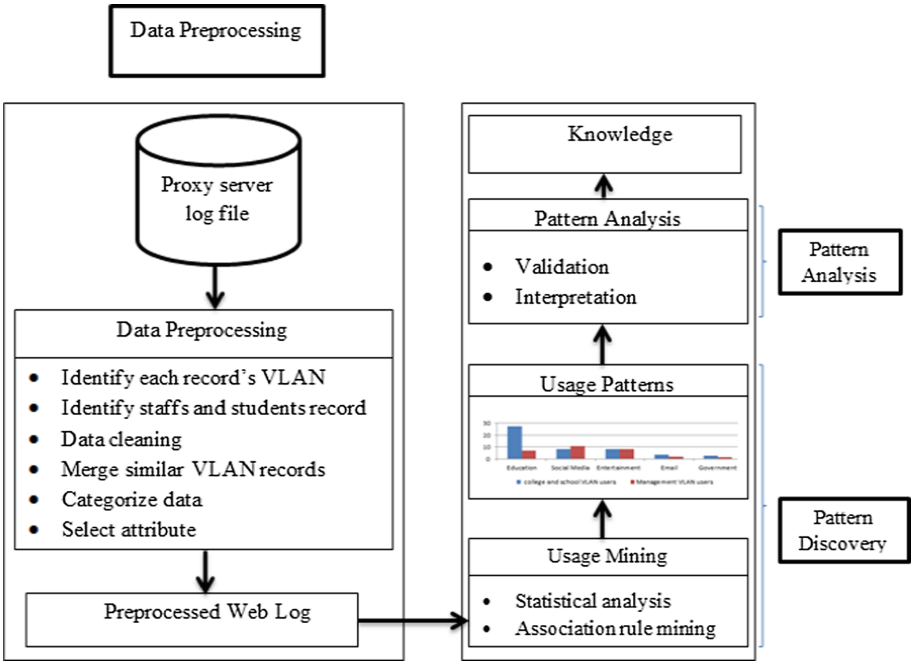

Fig. 1. Web usage mining process.

3 Result and Discussion

Six experiments are conducted. The first two experiments are statistical analysis and the remaining four are association rule mining.

3.1 Statistical Analysis

Experiment I: On Staffs Web Log Dataset: In this case a total of 60,019 data is used. As represented in Fig. 2, the x-axis represents web users VLAN and the y-axis represents the access frequency. Accordingly, two results are observed. Firstly, in some of the VLANs educational sites and in some others social media and entertainment sites accessed more frequently. In VLANs 80, 81 and 110, for instance, educational tutorials are accessed more frequently in the first priority. YouTube and Facebook are accessed next to educational tutorial sites. This is because VLANs 80, 81 and 110 are assigned to academic staffs. On the other hand, in the remaining VLANs (assigned for administrative staffs), Facebook is accessed more frequently in the first priority. The implication is that the interest of academic staffs is mostly accessing educational sites. While administrative staffs mostly accessed social media and entertainment sites. Secondly, the web traffic is analyzed based on number of query and download types. In the same VLANs, the web traffics is high since so many queries are submitted. As the

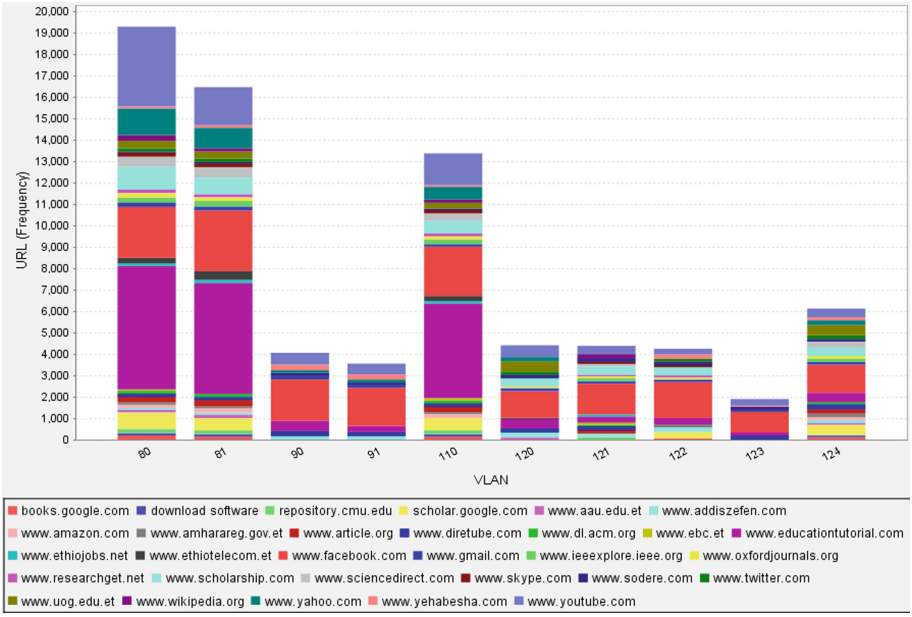

Fig. 2. Top frequently accessed sits and web traffic in the staffs VLANs.

number of query increases in the given bandwidth, the response time of the server decreases. Additionally, as the file content increase, such as video and audio and more people are accessing them, it consumes more bandwidth. This results more web traffic. Thus, for allocating bandwidths, knowing web traffic, access priorities and organizations mission are quite important.

Experiment II: On Students Web Log Dataset: In this experiment, a total of 110,802 data is used. As it has been shown in Fig. 3, in different VLANs educational, social media and entertainment sites were accessed. In VLANs 10, 11, 12, 13, 50, 73 and 74 Facebook is accessed more frequently, followed by educational tutorials and YouTube. As pointed out by domain experts, the previously specified VLANs are school (class room) computer laboratories. In VLANs 20, 21, 30, 41 and 42 educational tutorials sites are accessed more frequently. Entertainment and email sites are accessed next to educational sites. VLANs 20, 21, 30, 41 and 42 are library computer laboratories. Since Facebook is prohibited in library, students Facebook access frequency is null. However, in school computer laboratories, Facebook is accessed more frequently in the first priority. In some of the VLANs the web traffic is high, particularly in school (class room) computer laboratories than in the library computer laboratories.

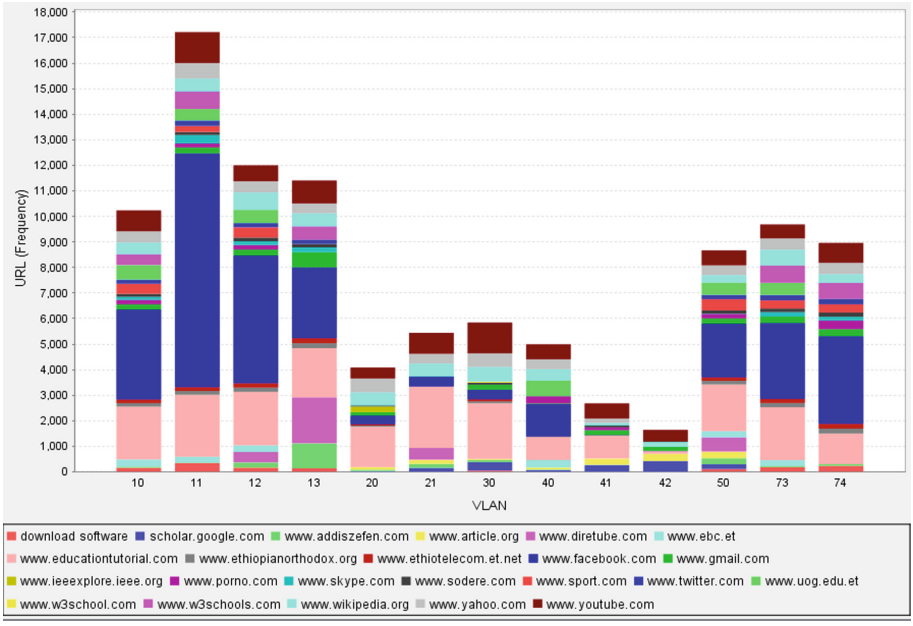

Fig. 3. Top frequently accessed sits and web traffic in the students' VLANs.

3.2 Browsing Behavior Characterization

In the previous experiments, we have shown that the web access behavior of staffs and students are different. This result is also supported by the statistical analysis as presented in Figs. 4 and 5. Accordingly, mostly the web browsing behavior of staffs is focus on educational sites in the first priority and then social media, entertainment and email sites, as depicted in Fig. 4. While students are interested in accessing social media in the first priority and then educational, entertainment and email sites, as depicted in Fig. 5.

There is also high web traffic in some school VLANs. This is due to high number of educational queries are submitted by the users in the first place. In administrative (management) VLANs social media and entertainment queries are submitted in the first and second place. In students VLANs, more web traffic is observed in schools and low web traffic is observed in library. The cause of more web traffic in schools are due to socila media, educational and entertainment queries, in first, second and third place, respectively. On the contrary, students accessed educational sites in library in the first place. Therefore, the web usage at UoG can be characterized as educational, social media, entertainment, email and government sites. The web traffic anaysis is made based on submitted query and file content type. Considering institutional mission, there are business critical activites and non-business activities. Obviously, in higher education, teaching and research activities can be considered as business critical activities. For higher educations social media and entertainment web traffics

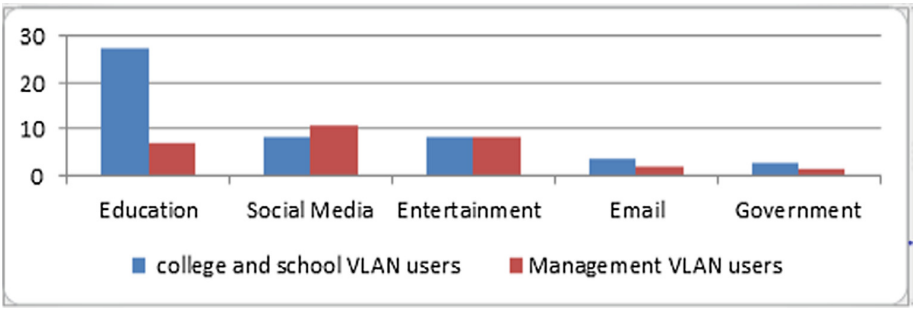

Fig. 4. Staffs' behavior characterization.

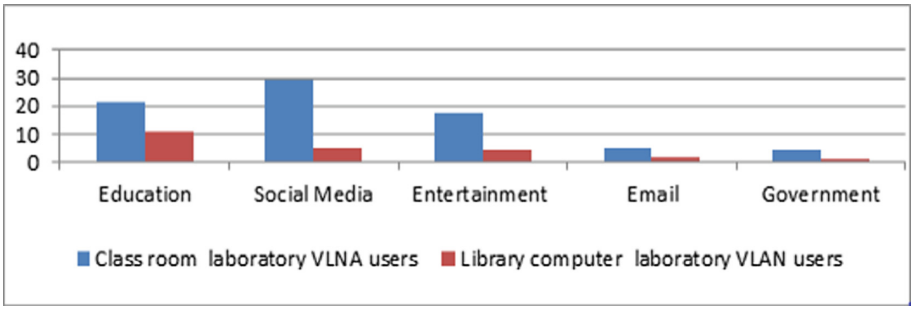

Fig. 5. Students' behavior characterization.

can be categorized under non-business related traffic. Therefore, allocating the required bandwidth to the business-critical traffic like is the primary solution for network performance improvement. Our findings can be used as an input for system performance and bandwidth management. As a result, this can be done using load balancing techniques.

3.3 Association Rule Discovery

We used objective (support and confidence) and subjective (domain experts) evaluations to determine interestingness of the rules [9].

Experiment III: Association Rule Discovery Using Apriori Algorithm on Staffs Web Logs: The value of min support and confidence was 0.01 and 0.9, respectively.

VLAN=81 URL24=accessed 4320 ==> URL35=accessed 4320 <conf:(1)> lift :(5.56) lev:(0.06) [3543] conv:(3543.58)

If URL24 accessed in VLAN 81 (academic staff), then URL35 had 100% probability to be accessed. This shows educational sites are accessed in VLAN 81,as presented in Experiment I.

Experiment IV: Association Rule Discovery Using FP Tree Algorithm on Staffs Web Logs: The value of min-support and confidence was 0.1 and 0.9, respectively.

```
[URL16=accessed, URL4=accessed]: 14382 ==> [URL1=accessed]: 14382   <conf:(1)> lift:(1.99) lev:(0.14) conv:(7167.78)
```

This rule states that if yahoo mail and educational tutorials browsed together, then Facebook had 100% probability to be accessed with them.

Experiment V: Association Rule Discovery Using Apriori Algorithm on Students Web Logs: The value of min-support and confidence was 0.01 and 0.9, respectively.

```
VLAN=11 URL8=accessed 3831 ==> URL1=accessed 3831   <conf:(1)> lift:(3) lev:(0.02) [2553] conv:(2553.72)
```

As of this rule, if URL8 accessed in VLAN 11, then Facebook had 100% probability to be access in the same VLAN. This shows entertainment and social media sites are accessed in VLAN 11 by the students, as presented, in Experiment II.

```
VLAN=30 URL4=accessed 1656 ==> URL20=accessed 1656   <conf:(1)> lift:(3) lev:(0.01) [1103] conv:(1103.88)
```

This rule shows that if educational tutorials accessed in VLAN 30, then URL20 had 100% probability to be access in this VLAN. Educational sites are accessed in library by the students, as presented in Experiment II.

Experiment VI: Association Rule Discovery Using FP Tree Algorithm on Students Web Logs: The value of min-support and confidence was 0.1 and 0.9, respectively.

```
[URL4=accessed, URL8=accessed]: 22635 ==> [URL1=accessed]: 22635   <conf:(1)> lift:(3) lev:(0.14) conv:(15088.37)
```

This rule states that if students browse educational tutorials and YouTube together, then they also accessed Facebook with 100% probability. To sum up, association rule discovered, social media and entertainment sites accessed more frequently in the students and administrative VLANs. On the contrary, educational and email sites accessed more frequently in academic staff VLANs.

4 Conclusion

The aim of this study was to understand web browsing behavior of UoG community. Statistical analysis and association rule mining algorithms are employed. The experimental results show that academic staffs focused on accessing educational sites and administrative staffs focused on social media and entertainment sites. On the other hand, it is also observed that students web browsing behavior in school and library are different. The identified usage patterns can be used for bandwidth management and system performance improvement.

References

1. Vellingiri, J., Pandian, S.C.: Survey on web usage mining. Glob. J. Comput. Sci. Technol. **11**(4) (2011)
2. Munilatha, R., Venkataramana, K.: A study on issues and techniques of web mining. Int. J. Comput. Sci. Mob. Comput. Mon. J. Comput. Sci. Inf. Technol. **3**(5) (2014)
3. Madhak, N.N., Kodina, T.M., Jayesh N. Varnagar, R.C.R.: Web usage mining using association rule mining on clustered data for pattern discovery. Int. J. Data Min. Tech. Appl. **2**(1) (2013)
4. Vijiyarani, S., Suganya, E.: Research issues in web mining. Int. J. Comput.-Aided Tech. (IJCAx) **2**(3), 55–64 (2015)
5. Santhosh Kumar, B., Rukmani, K.V.: Implementation of web usage mining using apriori and FP growth algorithms. J. Adv. Netw. Appl. **1**(6), 400–404 (2010)
6. Oskouei, R.J.: Identifying students behaviors related to internet usage patterns. IEEE Computer Science and Engineering Department Motilal Nehru National Institute Of Technology Allahabad (2010)
7. Amutha, K., Devapriya, M.: Web mining: a survey paper. Int. J. Comput. Trends Technol. (IJCTT) **4**(9), 3038–3042 (2013)
8. Uma Maheswari, B., Sumathi, P.: A comparative study of rule mining based web usage mining algorithms. Int. J. Sci. Res. (IJSR) **4**(11), 2540–2543 (2015)
9. Parvatikar, S., Joshi, B.: Analysis of user behavior through web usage mining. Int. J. Comput. Appl. (09750–8887) 27–31 (2014)
10. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. 2nd edn, pp. 243–246. Elsevier, Amsterdam (2006)
11. Kumar, R., Tomkins, A.: A characterization of online browsing behavior. In: International World Wide Web Conference Committee (IW3C2), 26–30 April 2010
12. Srivastava, J., Cooley, R.: Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor. **1**(2), 12 (2000)