Internet Activity Analysis Through Proxy Log

Kartik Bommeapally, Glisa T. K., Jeena J. Prakash, Sanasam Ranbir Singh and Hema A Murthy
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600036, India
Email: kartikaditya@gmail.com, {glisa, jeena, ranbir}@lantana.tenet.res.in, {hema}@cse.iitm.ac.in

Abstract—The availability of the Internet at the click of a mouse brings with it a host of new problems. Although the World Wide Web was first started by physicists at CERN to enable collation and exchange of data, today, it is used for a wide range of applications. The requirements on bandwidth for each of the applications is also varied. An Internet Service Provider must ensure satisfaction across the entire spectrum of users. To ensure this, analysis of Internet usage becomes essential. Further, an administrator can keep a record of user’s Internet activity and prevent unethical activities, since the Internet is also an excellent resource for providing anonymity. This analysis can also help in resource provisioning and monitoring. In this work, a web-based tool is first proposed to analyse the Internet activity. Next, data is collected from a proxy server at a campus-wide network. Traffic patterns of different types of users are studied. Finally, the paper concludes with strategies for monitoring and control of traffic.

I. INTRODUCTION

With increase in awareness and availability of Internet, information of any kind has become available at the click of a mouse. An administrator has to ensure that all the users get a fair share of the bandwidth. Today, addiction to the Internet is a serious issue amongst users, especially in campus-wide networks where Internet is freely available. An administrator in a campus may wish to ensure controls on Internet usage. In general, most campuses end up restricting usage over specific periods of the day. This has the disadvantage that a genuine user who needs access to information is also denied.

Internet traffic data can be collected from various sources such as routers, gateways or proxy servers. In this paper, we analyse a large proxy log to study “user access patterns”. Such a study can assist a network management system in traffic shaping and monitoring thus removing the necessity of regimentation. This is the main motivation of this paper. There are studies on the analysis of proxy log [1], [2], [3]. However, we focus on the following features:

• To keep machines behind it anonymous mainly for security.
• To speed up access to a resource (via caching).
• Contentfiltering through predefined rules.
• Logging Internet traffic.

II. PROXY-BASED INTERNET ACCESS

A proxy server acts as a go-between for requests from clients seeking resources from other servers. It evaluates every request according to its filtering rules and provides the resource by connecting to the relevant server and requesting the service on behalf of the client.

There are several functionalities of a proxy server ([4]). However, we focus on the following features:

• To keep machines behind it anonymous mainly for security.
• To speed up access to a resource (via caching).
• Contentfiltering through predefined rules.
• Logging Internet traffic.

III. DESIGN AND IMPLEMENTATION

In this section we briefly discuss the design and implementation of a proxy analyser. It has three major components namely log parser, database loader and data analyser. The basic framework is shown in Figure 1.

A. Log Parser

The purpose of this module is to extract useful information from the logs. The logs usually include the IP address and/or host name, the time of request, the user’s id, the URL requested, the status of the request. This module parses the logs and extracts the above information.
B. Database Loader

This module indexes the information obtained from log parser into the database. It has mainly four components. First component keeps track of the information such as user’s id and IP address. Second component keeps track of the domain information. In this study, we record only domain name, instead of entire URL. Third component keeps track of the access time. The information is stored in terms of the four quarters of a day. Mainly, this component keeps track of the time that a user spends on the Internet. However, proxy server logs only the time when a request has been made, not the time spent on Internet. In this study we use the following formula to compute the access time.

\[
 t_{\text{total}} = \begin{cases} 
 t_{\text{total}} + \Theta_{\text{cost}} & \text{if } t_{\text{cur}} - t_{\text{last}} > \Theta_{\text{limit}} \\
 t_{\text{total}} + (t_{\text{cur}} - t_{\text{last}}) & \text{Otherwise} 
\end{cases}
\]

where \(\Theta_{\text{limit}}\) is the maximum allowed time difference for two consecutive accesses, \(\Theta_{\text{cost}}\) is a system defined fixed value, \(t_{\text{cur}}\) is the current access time and \(t_{\text{last}}\) is the time that the user access last. If the time difference is above the threshold \(\Theta_{\text{limit}}\), only \(\Theta_{\text{cost}}\) is counted against the given user. After every estimation of the access time, \(t_{\text{last}}\) is updated with \(t_{\text{cur}}\) and estimation is repeated for all four quarters. The initial value of \(t_{\text{total}}\) is set to zero. The last component stores the relationship between the above three components. All these components are implemented satisfying all the integrity constraints and agreed upon by BCNF [5].

C. Data Analyser

Data analyser is the module which interacts with the user (i.e., network administrator). It provides the facility to view the statistics in the form of a graph. The analysis is done in two forms – offline and online. In offline analysis, the log data is collected and loaded into the system offline. However, in online analysis, log information is collected at run time. Whenever a new request arrives at the proxy server, it is automatically inserted into the database. Online analysis may slow down the activities of the proxy servers. It is advisable to deploy online analysis in networks with low traffic.

IV. ACTIVITY ANALYSIS

In this section we analyse the Internet usage pattern using the above tool. The analysis is made completely user centric for the administrator to keep track of user’s activities. Using the tool one can obtain an estimate of the activity over various websites accessed by the user. However we focus more on social-community websites in this study. The analysis presented in this section is broadly divided into three parts. First, we analyse the access pattern in terms of the amount of time that a user spends on Internet using the above tool. Second, we investigate traffic pattern in terms of the number of URLs requested by the users. Third, we propose few access control mechanisms to restrict users’ access without compromising the quality of service and investigate their effects on traffic sharing. Experiments reported in this paper are conducted on an Intel Core2 Duo machine with 1GB RAM.

A. Dataset

We first describe the characteristics of the proxy log used. For this study, we have used large proxy log collected from IITM 4 proxy server (running Squid [6]) over 5 months. IITM network has two proxy servers – Acad-Proxy and Hostel-Proxy. The Characteristics of the log files are shown in Table I.

B. Average Access Time Analysis

Various forms of analysis can be carried out using the tool discussed in Section III. However from an academic perspective, the reported analysis focuses on the following issues: (i) individual statistics: analysis over individual website accessed by the user, (ii) relative statistics: analysis over collective users for collective sites, and (iii) general statistics: analysis of access pattern by the category.

1. Individual Statistics: The proposed log analyser has provided a facility to explore the statistics of the user’s access time for a particular Web site. In Figure 2, we plot the average access time of a batch of users viewing the URLs matching a keyword “orkut”. Each batch is identified by a regular expression such as ’[A-Z][A-Z][0-9][0-9][A-Z][0-9][0-9][0-9][0-9]’. It compares the average access time for different quarters of a day over a period of time. This experiment has been conducted over different batches of users such as ’CS0[0-9][0-9][0-9][0-9],[CS0[0-9][0-9][0-9][0-9]’, ’CS0[0-9][0-9][0-9][0-9]’ and ’CS0[0-9][0-9][0-9][0-9]’. In Table II, we summarise the average access time for each batch defined above.

\(\Theta\) is set to 60min and \(\Theta_{\text{cost}}\) is set to 8min

---

1http://www.iitm.ac.in
2Q0: first quarter, Q1: second quarter, Q2: third quarter, Q3: forth quarter

---

TABLE I

<table>
<thead>
<tr>
<th>Characteristics of Proxy Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Duration</td>
</tr>
<tr>
<td>No. of request</td>
</tr>
</tbody>
</table>

Fig. 2. Average access time viewing the URL matching “orkut” for the set of students matching the regexp ’CS04b0[0-9][0-9]’ for the month of November 2008.
for the URLs matching “orkut”. It also shows that for the btech and dual degree students, the average access value is close to the maximum access value. The average access values across the four different programs at IITM show that they can be ordered as btech+dual > ms > phd > mtech.

2) Relative Statistics: In the relative analysis, we can compare the average access time across different users. In Figure 3, we explore the access time for the users matching the regular expression 'CS0[0-9][S0-9][0-9][0-9]' for the URLs matching orkut, faceook and ititm. This information can be used to perform inter/intra batch user’s access pattern analysis. We further perform various experiments across different batches of students. One interesting observation is the senior students have high Internet activity compared to their junior students.

3) General Statistics: In general statistical analysis, we explore the average access time of a user across different sites. To aggregate the results, we group the statistics by the category of the URLs that the user explored. This tool explores only certain number of URLs for each category. The sites in each category are manually selected from the list of popular sites.

Figure 4 shows the distribution of the access pattern of a user over the four domains for the month of December. The data is collected from Hostel-Proxy.

From the above analysis, it is clear that the distribution of average access time across the users is not uniform. There are students who spend most of their working hours browsing the Internet. It is important for an administrator to monitor every quarter the activity goes by the order browse > social-community > chat > academics. The periods 5-12-2008 to 8-12-2008 and 19-12-2008 to 27-12-2008 have 0 activity in all the quarters, which suggest that the user is on vacation and from 28-12-2008 onwards the activity again rises to high values in all the quarters. The access values are as high as 05:59:44 for browse, 05:47:58 for social-community, 05:48:49 for chat and finally 03:03:51 for acads during Q3 quarter (all the extreme values are recorded between 28-12-2008 and 31-12-2008). Tables III shows the number of connections to the server and the time spent on four different websites. Of all the websites google shows an alarming result of about 56329 connections having access time more than 1Hr for Q3, followed by orkut with 24009 connections made in Q3 having access time for more than 1Hr.

C. Traffic Analysis through URL Count

From the above analysis, it is clear that the distribution of average access time across the users is not uniform. There are students who spend most of their working hours browsing the Internet. It is important for an administrator to monitor
traffic at different instances in time and identify the users who are causing maximum traffic. In this section, we further investigate the number of URLs requested by the users over a period of time. URL count does not provide the actual network bandwidth consumed by the users. To estimate the actual bandwidth, we also need the information such as data transfer rate, size of the documents downloaded etc [7], [8]. Such information is not available in our dataset. However, the number of URLs reflects activities of a user on Internet. Therefore, we can use the URL count as a measure to approximate the traffic. The simple hypothesis is that larger the number of URLs requested, higher is the Internet traffic caused. Such analysis helps the administrator to perform various Internet traffic shaping based on (i) the type of URLs that the user visited, (ii) the usage of restricted URLs, (iii) restricting certain users from accessing certain URLs at the time when there is slow connection or busy traffic. The analysis reported in this section is done independent of the tool discussed in Section III.

1) Users’ Classification: Figure 5 shows the distribution of the average number of URLs requested by the users per day. It clearly shows that majority of the users have a small average number of URLs accessed per day and few users have extremely large number of requests.

Based on the average number of requests generated by the users, we further classify the users into three band—low, middle and high. The bound of each band is defined using the following expressions

\[ l = \mu - c \cdot \sigma \]  \hspace{1cm} (1)

and

\[ u = \mu + c \cdot \sigma \]  \hspace{1cm} (2)

where \( \mu \) is the average number of URLs generated by a user in a day, \( \sigma \) is the standard deviation and \( c \) is a constant in [0, 1]. Users with average count less than \( l \) are placed in the low band, between \( l \) to \( u \) are placed in middle band and greater than \( u \) in high band. Table IV shows the size of each band. It clearly shows that number of users in the high band is smaller compared to other classes.

In Figure 6, we investigate the percentage of traffic (i.e., URL counts) generated by the users in each band. It clearly shows that though the number of users in high band is small, it generates a considerable amount of traffic i.e., almost 45%.

In Figure 7, we show the percentage of the traffic generated by the users at different time intervals of a day. It clearly shows that the distribution is not uniform. The traffic during fourth quarter is higher than the other quarters.

It is also important to investigate the traffic caused by certain Web sites. In Figure 8, we plot the percentage of URLs containing popular keywords. The keywords are selected using term frequency. It is clearly observed that a significant portion of the traffic is caused by the URLs containing the keywords such as “orkut”, “facebook”, “talk”, “youtube” which are discouraged in many organisations.

Entropy is a popular measure used to analyse the distribution of the symbols of a random variable. If the entropy is high, then the distribution is uniform and if entropy is small then the distribution is skewed. We also apply entropy to study the distribution of URLs visited by the users. A user with a small entropy indicates that the user confines his/her activities mostly to few domains. Such a study can also be used to analyse user specific traffic shaping to decide whether a particular user needs to be granted more access to certain sites or to restrict access to certain sites. We use normalised entropy to study this distribution.

\[ H = -\frac{\sum_{u \in U} P(u) \cdot \log(P(u))}{\log(|U|)} \]  \hspace{1cm} (3)

where \( U \) is the list of URLs access by a user. In Figure 9, we show the distribution of the entropy using the distribution of the most popular 50 URLs. It clearly shows that there are users who mostly confine their activities to only very few URLs.

D. Traffic Control

In the above discussion, we explore users’ Internet access pattern. In this section, few access control mechanisms are

\[ \text{SIZE OF EACH USER’S CLASS FOR THE MONTH OF FEB. 2009} \]

\begin{tabular}{|c|c|c|}
\hline
User’s type & lower band & middle band & higher band \\
\hline
\text{Users} & 862 & 2684 & 302 \\
\hline
\end{tabular}
discussed to investigate feasibility of controlling users’ Internet activity through proxy servers. To analyse the effects of these mechanisms, we have run simulation programs over the original dataset. Figure 10 shows the distribution of the URLs at different time intervals of a day after applying access control mechanisms. There are several ways to control users’ Internet access. However we have focused on the following.

1) First, we apply restriction on the number of URLs requested by the users in a day (limits to average number of URLs per day). Figure 10 clearly shows that the traffic reduces significantly (almost 56%) after applying this control mechanism. However this mechanism is not flexible. It may results heavy traffic at one time and no traffic at another time of a day.

2) Second, we apply restriction on the domain names. We have ignored all the URLs containing the keywords “megaupload”, “orkut”, “talkgadget”, “voice”, “facebook”. The traffic reduces significantly (almost 34%) after ignoring the URLs. From the figure, it clearly indicates that users are likely to access these social networking sites during evening and night. The reduction at different quarters can be ordered as Q3 > Q2 > Q0 > Q1.

3) The above restrictions, sometimes, may cause denial of important information access. A better control mechanism will be to restrict dynamically based on traffic load. Lastly, we apply few access control mechanisms and investigate their effects on Internet traffic.

V. CONCLUSION

In this paper, we discuss a proxy log analyser. Using this tool, we analyse the amount of time that a user spends on Internet. We further investigate the traffic generated by users by exploring URL count. Lastly, we apply few access control mechanisms and investigate their effects on Internet traffic.

REFERENCES