SquiGIS: A web-based Geographic Information System for the UPLB Network

Shiela Kathleen L. Borja , Ludwig Johann B. Tirazona , and Joseph Anthony C. Hermocilla

***Abstract*— The researchers developed a system which resides on a machine different from the one keeping the Squid access logs. Both machines belong to the same network. This was done to allow larger memory allocation for the system. The researchers created a log parser to get the logs from the machine containing the Squid access logs through SSH every two hours. The parsed data is stored in a database managed by squiGIS. The data stored in the database was used to analyze the logs and create a web-based monitoring module for the system. Search functionalities specified by the UPLB Information Technology Center were also incorporated into the system.**

***Index Terms*— Geographic Information System, network monitoring, squid log parser and analyzer**

I. Introduction

A. Background of the Study

The University of the Philippines Los Baños(UPLB) network has a very essential role in rendering the university's services to the students, faculties, and staffs effectively. Thus, it is very important to maintain the network's integrity and performance through an effective monitoring system. Logs are files maintained by the proxy servers to keep track of the data about the users and their Internet activities. Data from the logs are extracted and used as input to the monitoring system to generate reports about the Internet activities of the users.

Currently, the UPLB Information and Technology Center (ITC) uses Squid 3.0 as the proxy server for the UPLB network. A proxy sever has two main functionalities in the network. The proxy server acts as an intermediary between the client and the server containing the requested resources and requests for the resources in behalf of the client. It can also be used to restrict access to undesired sites.

The objective of this project was to build a monitoring system for the Internet usage in the UPLB network through processing the logs generated by the Squid proxy server used in the UPLB network. Then, the GIS was used to help monitor the Internet activities in the network at a certain time.

B. Statement of the Problem

ITC uses Squint, Calamaris, and LightSquid as log parsers but no single log parser can suffice the requirements that ITC needs to monitor the activities in the UPLB network.

The Squint log parser is an open-source Squid log file analyzer which produces static HTML reports. [1] Calamaris is a log parser for Squid, NetCache, Inktomi Traffic Server, and other proxy servers and generates reports including Peak-usage, Request Methods, Status Report of incoming and outgoing requests, second and Top-level destinations, content-types and performance. [2] [3] LightSquid is a perl-based cgi Squid logfile parser. [1]

ITC uses the all of these three log parsers in order to combine the good features of each log parser and analyze the Internet activities in the network. Squint is used more often because it is faster than LightSquid but Squint is also inefficient because it eats up a lot of CPU time [4]. Calamaris analyzes the overall statistics of the logs but it does not analyze individual information of each client.

Although Squint, Calamaris, and LightSquid are powerful tools, UPLB ITC needed a more customized log file analyzer to better suit the unit's monitoring transactions.

C. Significance of the Study

This study aimed to help improve monitoring of the Internet resources through adding the necessary functionalities as required by the UPLB ITC. Improved monitoring of the Internet activities in the network would help track violators. Logs and reports about the internet activities of the violator could serve as a hard evidence against the violator and could also serve as a basis for the proper sanctions that would be given to the violator. [5]

This study also aimed to help ITC to use only one log parser which could cater all their requirements since the specifications of the system depend on their requirements. Additional features such as the list of *View by Top Users According to bandwidth consumed, Top IP Addresses According to bandwidth consumed, Most Accessed Domain, View by IP Address, View by Lightweight Directory Access Protocol (LDAP) User name, View by Subnetwork, View by Active Hosts, Top IP Addresses with most Access Denies* , and *Show by Date* were included to improve the searching functionality of the system. The locator could also help in visualizing the Internet activities in the network since the data about the number of requests and total bandwidth for each *Active Host* were plotted on the UPLB map. Using the locator, the administrator could have a quick look on the distribution of the client requests during a certain time of the day.

D. Scope and Delimitation

The system was limited to mapping wired connections, that is excluding clients connected through wireless connection

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

such as the campus WiFi, but activities of clients connected through wireless connection was still monitored. The system extracted the LDAP user name of the user from the access log file, if it exists, then the LDAP user name and IP Address was used to represent a user. If LDAP username was unavailable, IP address was used to represent a user. The system only used the squid log format, *%ts.%03tu %6tr %⟩a %Ss/%03Hs %⟨st %rm %ru %un %Sh/%⟨A %mt* where *ts* is the seconds since epoch, *tu* is the subsecond time, *tr* is the response time,⟩*a* is the client source IP address, *Ss is the Squid request status*, *Hs* is the HTTP status code,⟨*st* is the reply size including the HTTP headers, *rm* is the request method,*ru* is the request URL,*un* is the username, *Sh* is the Squid hierarchy status, ⟨*A* is the Server IP address or peer name and *mt* is the MIME content type.

E. Objectives

The main objective of this study was to develop a monitoring system for the UPLB network.

Specifically, the study aimed to:

1) Create a Squid log parser using PHP 5.0 to access the Squid access log file.
2) Create a monitoring module which directly accesses the log file and stores parsed data into the database.
3) Create a web-based user interface for the monitoring system.
4) Add search functionalities as specified by the UPLB ITC.

F. Date and Place of Study

This study was conducted at the Institute of Computer Science, College of Arts and Sciences, University of the Philippines Los Bañosfrom October 2011 to February 2011.

II. Related Work

A number of Log Parsers were already available for the public. Majority of the log parsers were scripts written in Perl, while a small number were developed as a web-based application. An example of a web-based log analysis software is Rotilia developed by Mehdi Adibi. Rotilia is a web-based log-file analyzer which was written in PHP. It outputs the total time the cache server was busy, total size of packages, and sites accessed. It can also show the results for every user at any given period of time. [6]

Bommepally et al. (2010) conducted an Internet Activity Analysis using a Proxy Log. In the said study, the researchers proposed a web-based tool to monitor the Internet activity and implemented it in a campus-wide network.They proposed a design and implementation of the log parser analyzer to be used in the study. The log parser analyzer they used has three main partslog parser, database loader, and data analyzer. The log parser is a module which parses the logs and extracts necessary information. The database loader is the module responsible for indexing the data extracted by the log parser from the log file. The data analyzer is the module used to interact with the user. [7]

III. Theoretical Framework

A. Squid Log Files

Squid log files were commonly used to determine Squid's workloads and performance. [8] Log files maintained in Squid contained data about the *access information, configuration errors*, and *resource consumption*. [8] If Squid can not write to its log files, it will shutdown itself that is why it is important to maintain them. Squid's rotate feature is used to properly manage the log files. [8]

Squid log file rotation is a very important concept used in managing Squid log files. Rotation in Squid is issued using the command *squid -k rotate*. Rotation takes place in a specified time of the day. Log files are compressed and renamed, then a new log file is created. By default, the number of rotated log files to be generated is 10, this means that Squid will keep 10 access logs before overwriting the oldest log. The *logfile_rotate* tag can be used to specify the preferred number of log files to be rotated. [9]

Three log files used by squid, the *access, cache*,and *store* logs. One of the frequently used log file among the three is the *access log* which is used by log file analysis programs. *Access log* contains basic access information such as time of access, response time, client source IP address, and other relevant information.

B. Parser

According to Barnbrook (2002), A parser is a "procedure" with two main functionalities: first is to recognize the sentence and the second is to determine how it is built. [10] According to Hoftcroft et al. (2001), a parser "discovers the structure of a program". [11]

In the case of a log parser, ideally, it recognizes the format to which the logs are written and use that information to extract the data from each log entry. But most of the time the log parser assumes a certain log format to which all the log files are commonly written. This log format will be used to parse the log files even if it is not written using the same format causing some discrepancies among the data extracted from the logs.

C. Geographic Information System (GIS)

Geographic Information System is a mix of geography, cartography, some database theory, and mathematics.

In this paper, the researchers considered the definition used by Clarke and Estes and Star. Clarke(1995) defined GIS as an "automated system for capture, storage, retrieval, analysis, and display of spatial data" [12] Star and Estes (1990), as cited by Clarke(1995), also defined GIS as "an information system that is designed to work with data referenced by spatial or geographic coordinates. In other words, a GIS is both a database system with specific capabilities for spatially-referenced data, as well as a set of operations for working with the data". [12] The data is spatial when one of the attributes of an object locates the object in a map. As described by Clarke (1995), information mapped in the GIS can be used to solve problems, perform queries, come up with the answer, or try a possible solution. [12]

IV. Methodology

A. Hardware Specification

1) Ubuntu Operating System (OS)

B. Software Interface

1) Squid 3.0
2) Apache 2
3) MySQL 5.1.41
4) PHP 5.0
5) JavaScript

C. Plug-in

1) JQuery UI
2) JQuery DataTables
3) Open Street Maps

D. Method

The system was deployed in a machine different from that of the machine keeping the *access.log* file which is rotated every two hours. After the *access.log* file was rotated, a *rotate.log* file was generated which was used by the system to indicate whether the logs already rotated.

The *Monitor Module* runs using cron, scheduled every two hours. The *Monitor Module* is composed of three parts– Access, Analyzer, and Parser. For the *Access* part, the system used SSH to copy the contents of the *access.log.0* file after the logs were rotated else it would clear the *outputfile.log* file. The *outputfile.log* where logs are copied from the Squid's *access.log.0* file and the current timestamp was stored in the database to represent the latest run of the system. The *Analyzer* part checked which log entries to consider, read each log entry, and checked if it was the last log entry read by squiGIS. The last log entry and timestamp of the last log entry were stored in the database. The *Parser* part parsed each line of log entries to get the data needed and stored it to the database maintained by the system. The system checked the subnetwork where the IP Address of the user belongs to, if the IP Address did not match any of the currently listed subnetwork, it was classified into the "OTHER LOCATION" subnetwork.

The *Locator Module* of the system also ran using cron, and was scheduled every two hours each time the system accesses new logs. The *Locator Module* got the *Active Hosts* or the nodes which made requests for the past two hours since the last run of the system. Data about the location, coordinates, number of requests, and total bandwidth consumed of the *active hosts* were determined. The *Active Hosts* were grouped according to their location and the data was supplied to the Map.

Every time the user performed queries using the *Display Module*, the *Display Module* accessed the *Analyzer Module* which in turn accessed data stored in the database to perform database queries. This module provided data to the *Display Module* to display as the query results.

The *Display Module* could be viewed using a browser. It used a jQuery plug-in called *dataTables* to present the query results to the user. This module had three general functionalities–*Add Location, View Report, and View Map of Active Hosts*. *Add Location* enabled the user to add, edit, and delete subnetworks, *View Report* allowed the user to use the search functionalities of the system, and the *View Map of Active Hosts* enabled the user to have a quick look on the traffic on a certain time of the day. The system was only made be accessible to UPLB ITC admin.

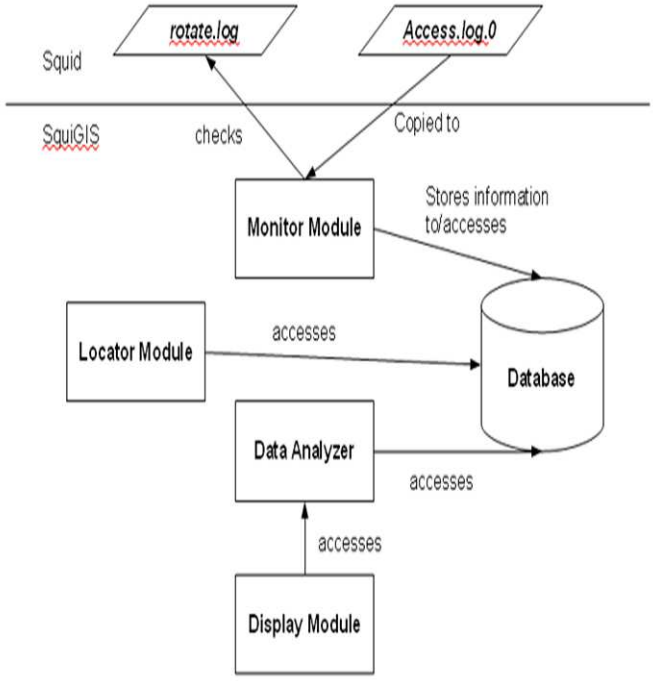


Fig. 1. SquiGIS Design

The requirements for the system included: *View by Top Users according to total bandwidth consumed, View Top IP Addresses according to total bandwidth consumed, Most Accessed Domain, View IP address, View User name, View by Subnetwork, View Active Hosts, Show By Date, Top IP Addresses with most Access Denies, Daily Summary, Weekly Summary, Monthly Summary, and Keep daily logs for a maximum period of 14 days*. [5]

The database design of squiGIS is shown in the Appendix.

V. Results and Discussion

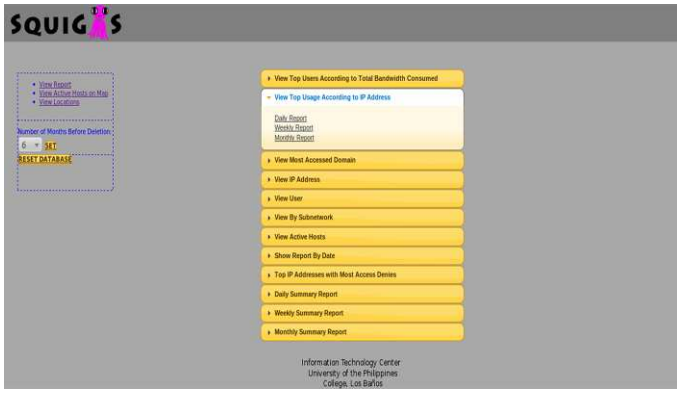


Fig. 2. Main Menu

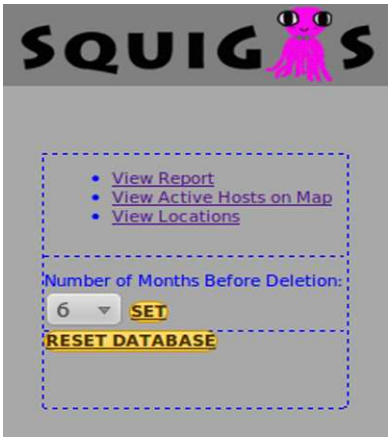


Fig. 3. Menu Panel

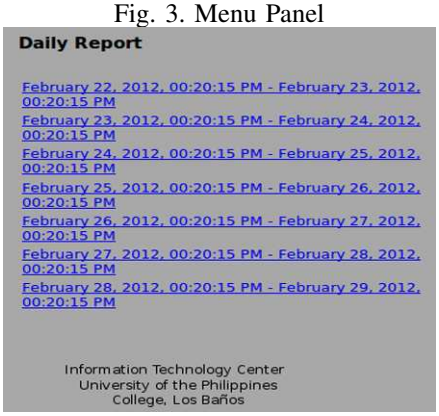


Fig. 4. Daily Report

Fig. 5. Weekly Report

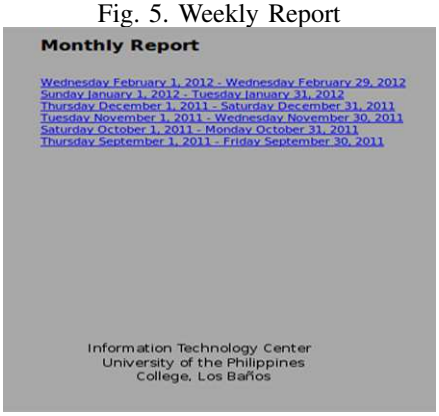


Fig. 6. Monthly Report

Fig. 7. View Top Users Menu

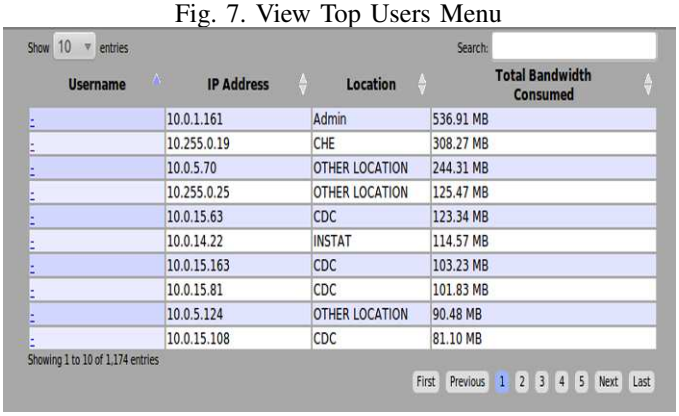


Fig. 8. View Top Users according to Bandwidth Consumed

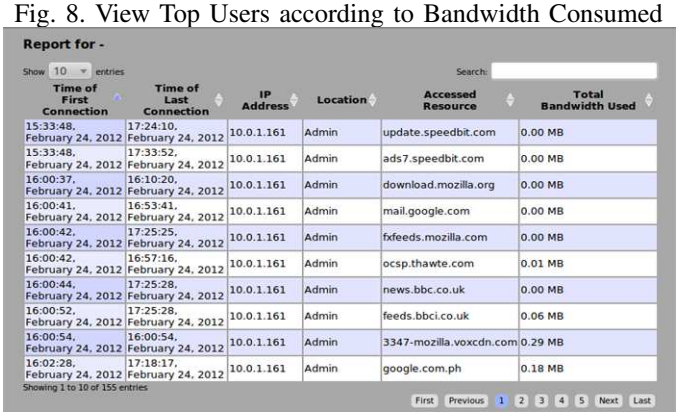


Fig. 9. View Top Users according to Bandwidth Consumed Detailed Report

Fig. 10. View Top Users according to Bandwidth Consumed Detailed Report

Fig. 11. View Top IP Addresses according to Bandwidth Consumed, Menu

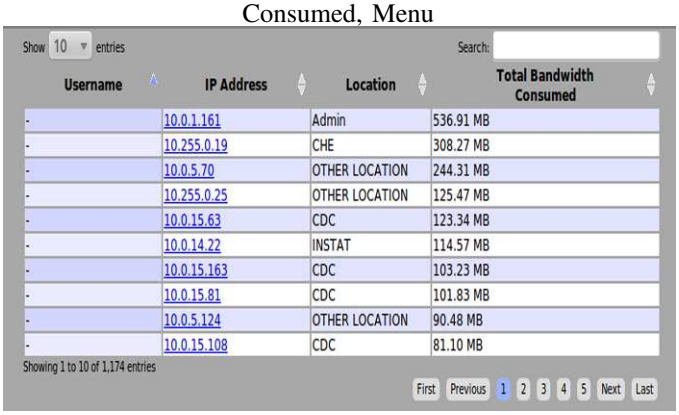


Fig. 12. View Top IP Addresses according to Bandwidth Consumed

Fig. 13. View Top IP Addresses according to Bandwidth Consumed Detailed Report

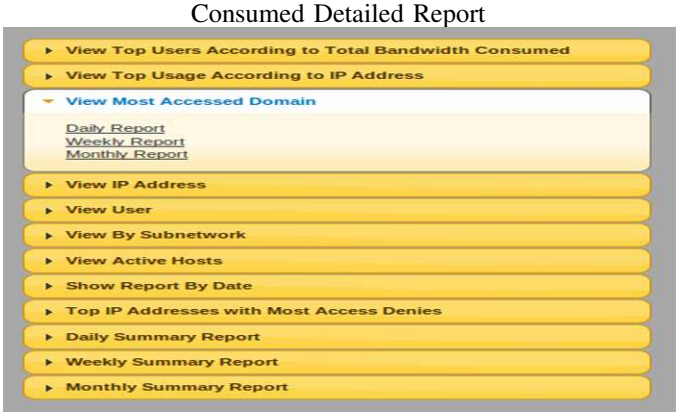


Fig. 14. View Most Accessed Domain, Menu

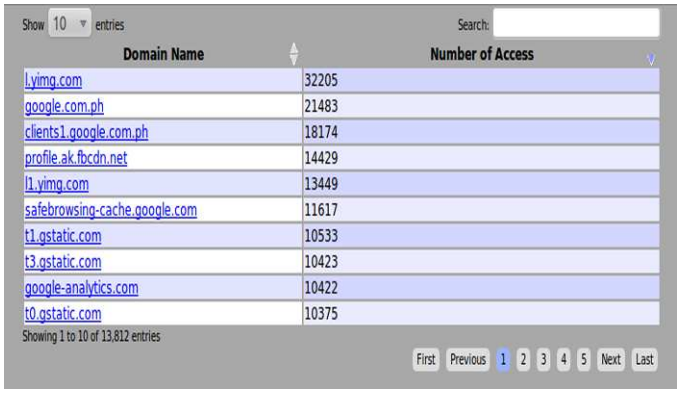


Fig. 15. View Most Accessed Domain

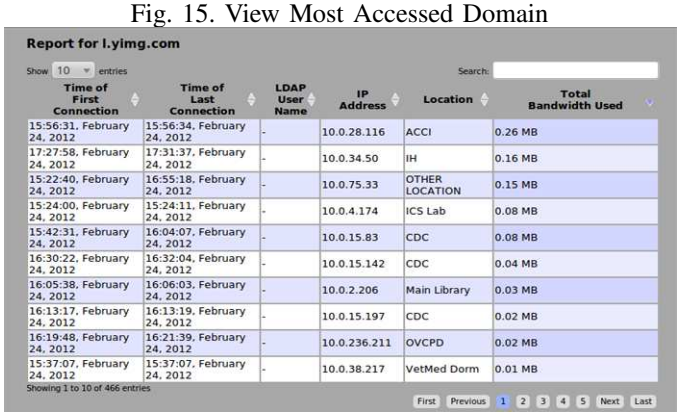


Fig. 16. View Most Accessed Domain Individual Report

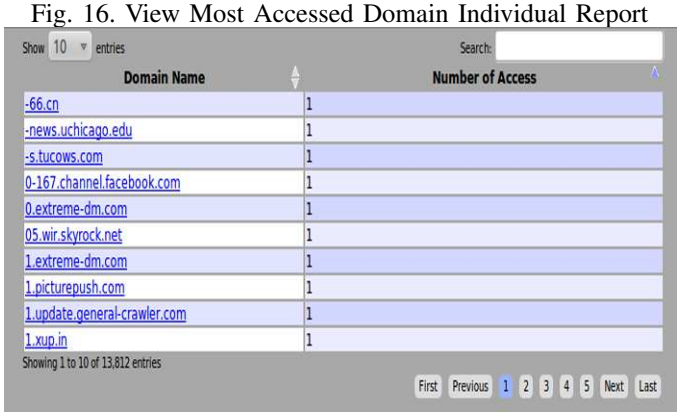


Fig. 17. View Least Accessed Domain

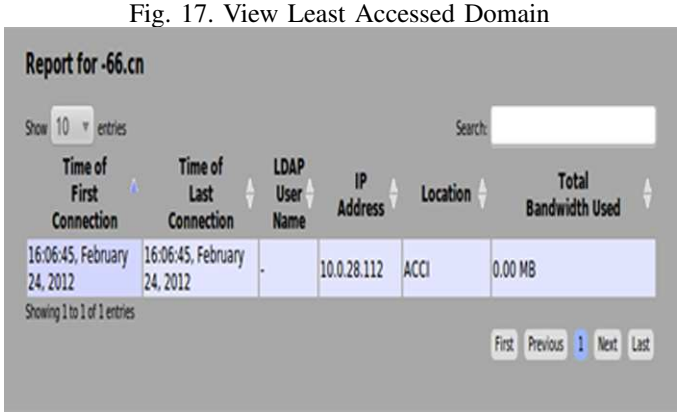


Fig. 18. View Least Accessed Domain Individual Report

Fig. 19. View IP Addresses, Menu

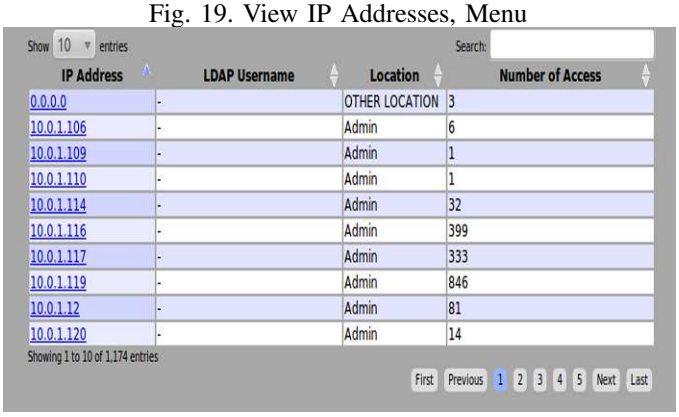


Fig. 20. View IP Addresses

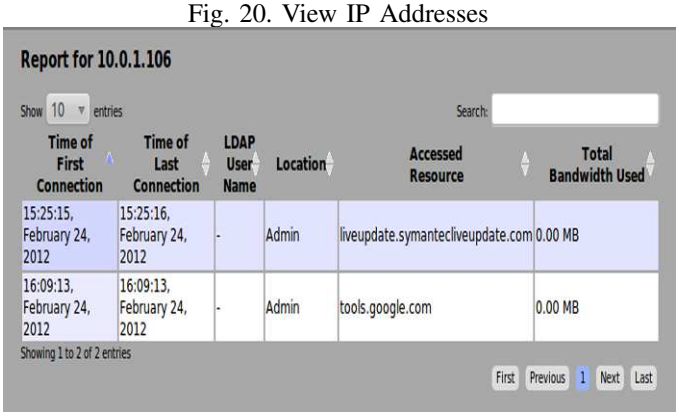


Fig. 21. View IP Addresses Individual Report

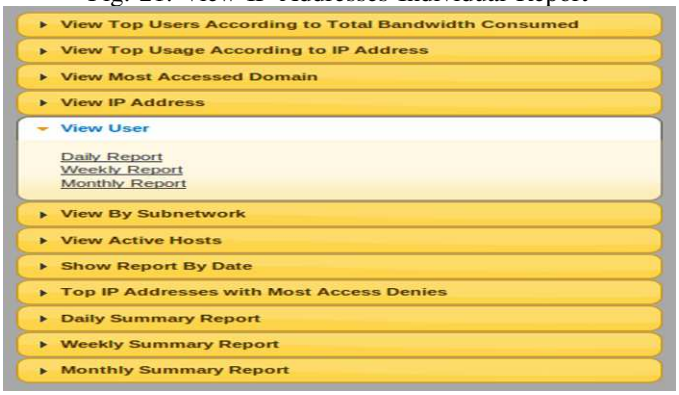


Fig. 22. View Users, Menu

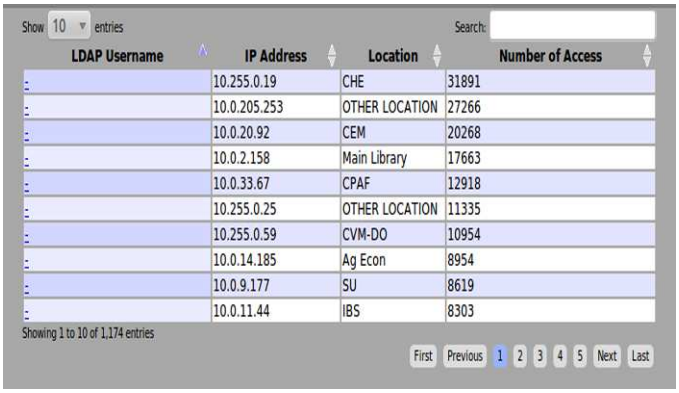


Fig. 23. View Users

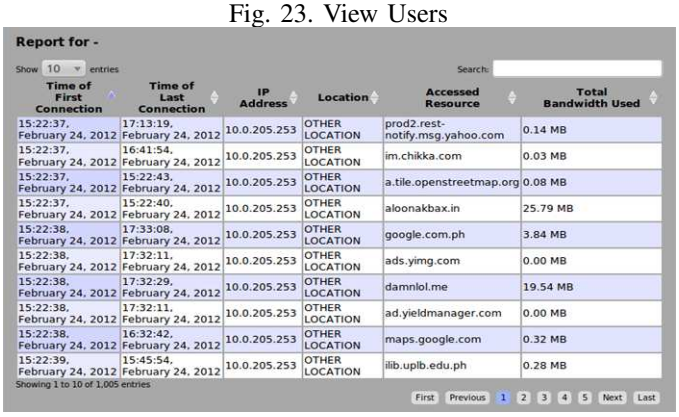


Fig. 24. View Users, Individual Report

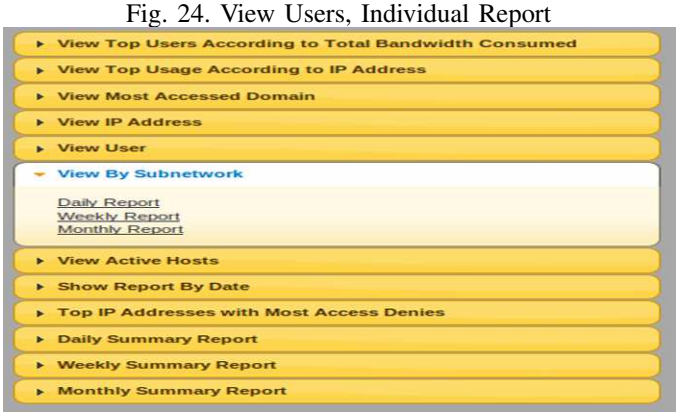


Fig. 25. View by Subnetwork, Menu

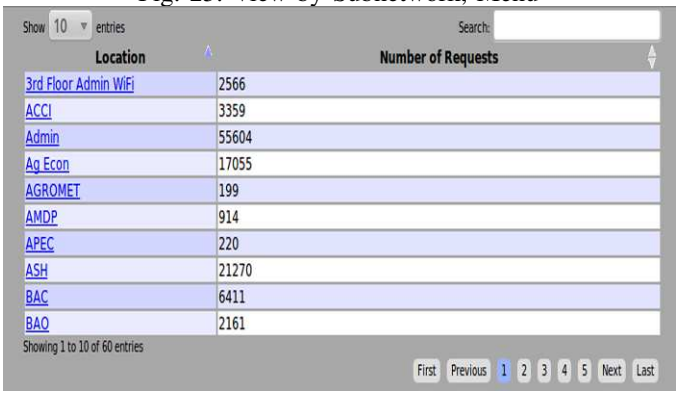


Fig. 26. View by Subnetwork

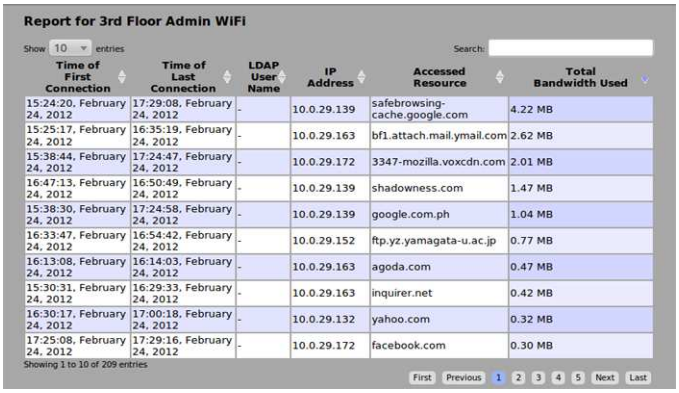


Fig. 27. View by Subnetwork, Individual Report

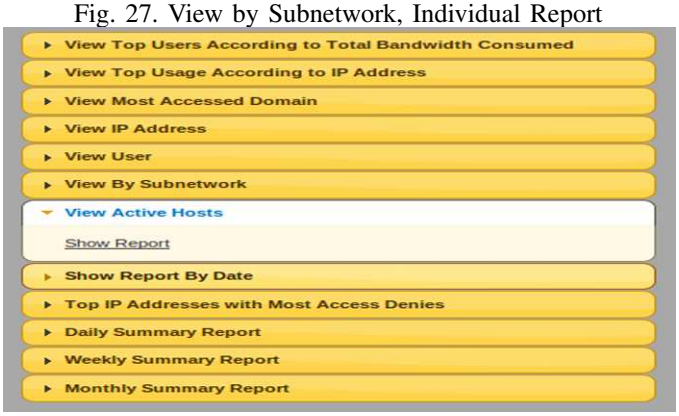


Fig. 28. View Active Hosts Menu

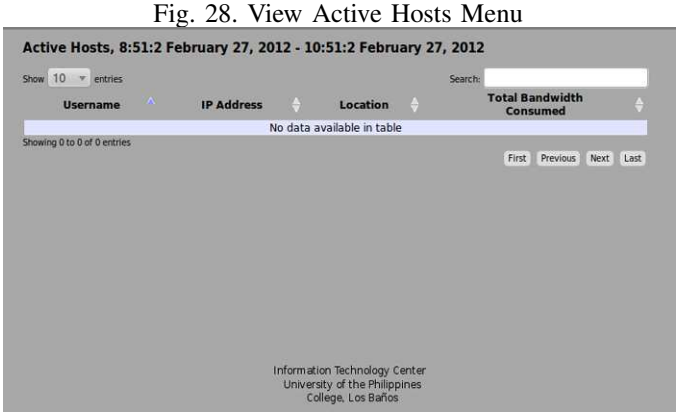


Fig. 29. View Active Hosts

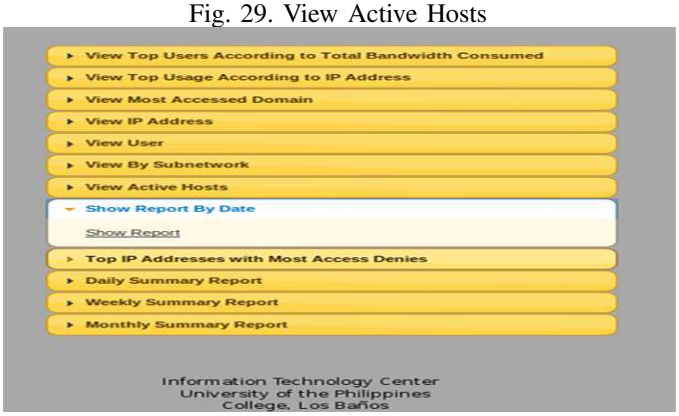


Fig. 30. Show by Date, Menu

Fig. 31. Show by Date, Calendar

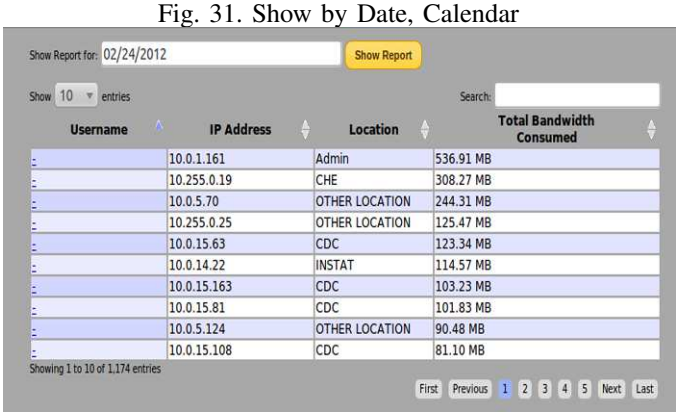


Fig. 32. Show by Date, Summarized Report

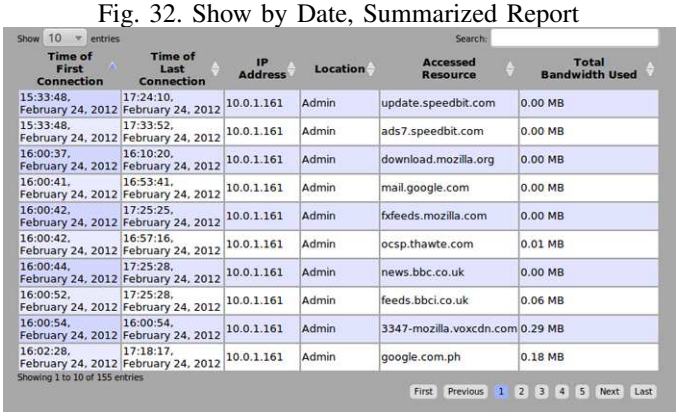


Fig. 33. Show by Date, Individual Report

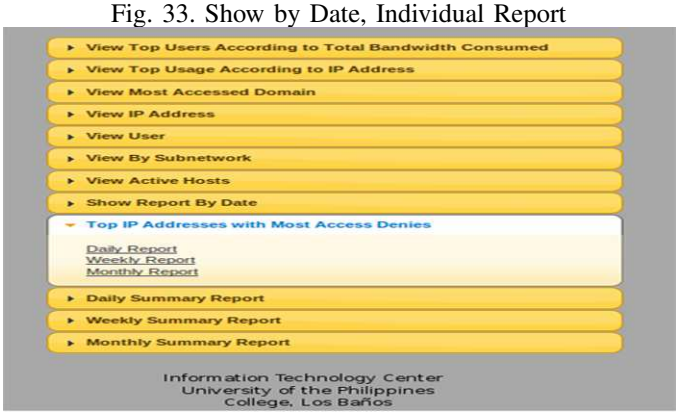
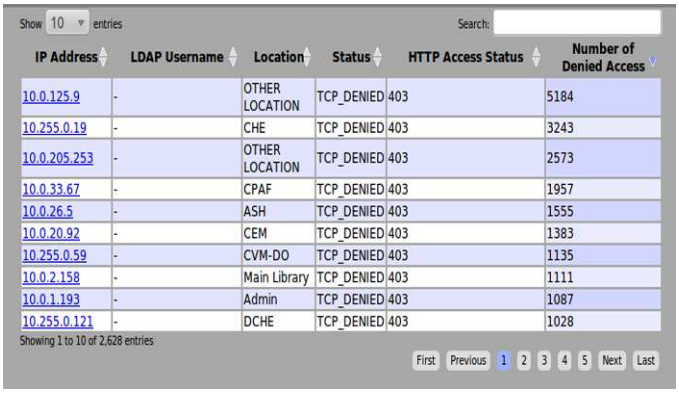


Fig. 34. View IP Addresses with Most Access Denies, Menu



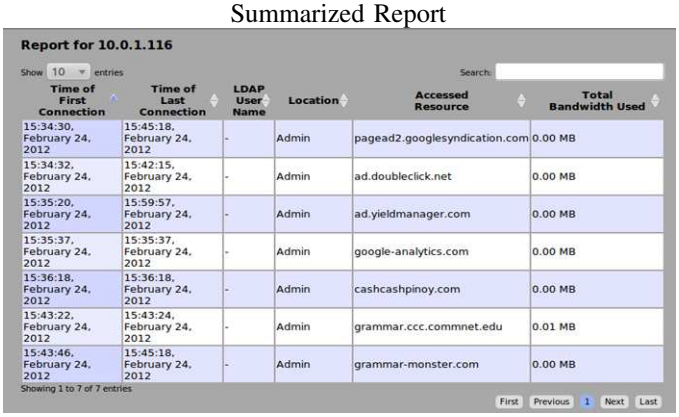
Show 10 entries Search:

IP Address	LDAP Username	Location	Status	HTTP Access Status	Number of Denied Access
10.0.125.9	-	OTHER LOCATION	TCP_DENIED	403	5184
10.255.0.19	-	CHE	TCP_DENIED	403	3243
10.0.205.253	-	OTHER LOCATION	TCP_DENIED	403	2573
10.0.33.67	-	CPAF	TCP_DENIED	403	1957
10.0.26.5	-	ASH	TCP_DENIED	403	1555
10.0.20.92	-	CEM	TCP_DENIED	403	1383
10.255.0.59	-	CVM-DO	TCP_DENIED	403	1135
10.0.2.158	-	Main Library	TCP_DENIED	403	1111
10.0.1.193	-	Admin	TCP_DENIED	403	1087
10.255.0.121	-	DCHE	TCP_DENIED	403	1028

Showing 1 to 10 of 2,628 entries

First Previous 1 2 3 4 5 Next Last

Fig. 35. View IP Addresses with Most Access Denies, Summarized Report



Report for 10.0.1.116

Show 10 entries Search:

Time of First Connection	Time of Last Connection	LDAP User Name	Location	Accessed Resource	Total Bandwidth Used
15:34:30, February 24, 2012	15:45:18, February 24, 2012	-	Admin	pagead2.googlesyndication.com	0.00 MB
15:34:32, February 24, 2012	15:42:15, February 24, 2012	-	Admin	ad.doubleclick.net	0.00 MB
15:35:20, February 24, 2012	15:59:57, February 24, 2012	-	Admin	ad.yieldmanager.com	0.00 MB
15:35:37, February 24, 2012	15:35:37, February 24, 2012	-	Admin	google-analytics.com	0.00 MB
15:36:18, February 24, 2012	15:36:18, February 24, 2012	-	Admin	cashcashpinoy.com	0.00 MB
15:43:22, February 24, 2012	15:43:24, February 24, 2012	-	Admin	grammar.ccc.commnet.edu	0.01 MB
15:43:46, February 24, 2012	15:45:18, February 24, 2012	-	Admin	grammar-monster.com	0.00 MB

Showing 1 to 7 of 7 entries

First Previous 1 Next Last

Fig. 36. View IP Addresses with Most Access Denies, Individual Report

Report for 10.0.125.9

Show 10 entries Search:

Time of First Connection	Time of Last Connection	LDAP User Name	Location	Accessed Resource	Total Bandwidth Used
16:26:09, February 24, 2012	16:46:44, February 24, 2012	-	OTHER LOCATION	img11.imageshack.us	0.00 MB
16:26:09, February 24, 2012	16:43:13, February 24, 2012	-	OTHER LOCATION	badge.facebook.com	0.00 MB
16:26:10, February 24, 2012	16:43:14, February 24, 2012	-	OTHER LOCATION	i974.photobucket.com	0.00 MB
16:26:14, February 24, 2012	16:43:15, February 24, 2012	-	OTHER LOCATION	facebook.com	0.00 MB
16:26:14, February 24, 2012	16:43:14, February 24, 2012	-	OTHER LOCATION	dg.specificclick.net	0.00 MB
16:26:15, February 24, 2012	16:43:14, February 24, 2012	-	OTHER LOCATION	static.ak.fbcdn.net	0.00 MB

Showing 11 to 16 of 16 entries

First Previous 1 2 Next Last

Fig. 37. View IP Addresses with Most Access Denies, Individual Report

Report for 10.0.125.9

Show 10 entries Search:

Time of First Connection	Time of Last Connection	LDAP User Name	Location	Accessed Resource	Total Bandwidth Used
15:30:50, February 24, 2012	16:56:02, February 24, 2012	-	OTHER LOCATION	ad.yieldmanager.com	0.00 MB
15:31:09, February 24, 2012	16:56:32, February 24, 2012	-	OTHER LOCATION	geo.yahoo.com	0.00 MB
15:39:13, February 24, 2012	16:06:51, February 24, 2012	-	OTHER LOCATION	ads.yimg.com	0.00 MB
15:42:48, February 24, 2012	15:42:48, February 24, 2012	-	OTHER LOCATION	dnl-05.geo.kaspersky.com	0.00 MB
16:26:09, February 24, 2012	16:43:20, February 24, 2012	-	OTHER LOCATION	google-analytics.com	0.00 MB
16:26:09, February 24, 2012	16:43:13, February 24, 2012	-	OTHER LOCATION	pagead2.googlesyndication.com	0.00 MB
16:26:09, February 24, 2012	16:46:45, February 24, 2012	-	OTHER LOCATION	img59.imageshack.us	0.00 MB
16:26:09, February 24, 2012	16:46:33, February 24, 2012	-	OTHER LOCATION	img3.imageshack.us	0.00 MB
16:26:09, February 24, 2012	16:46:22, February 24, 2012	-	OTHER LOCATION	img33.imageshack.us	0.00 MB
16:26:09, February 24, 2012	16:46:41, February 24, 2012	-	OTHER LOCATION	img684.imageshack.us	0.00 MB

Fig. 38. View IP Addresses with Most Access Denies, Individual Report

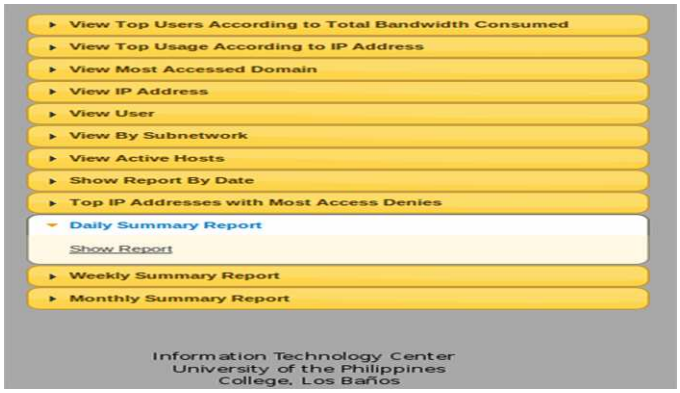
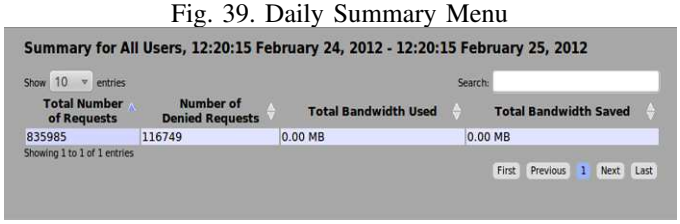


Fig. 39. Daily Summary Menu



Summary for All Users, 12:20:15 February 24, 2012 - 12:20:15 February 25, 2012

Show 10 entries Search:

Total Number of Requests	Number of Denied Requests	Total Bandwidth Used	Total Bandwidth Saved
835985	116749	0.00 MB	0.00 MB

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Fig. 40. Daily Summary Report

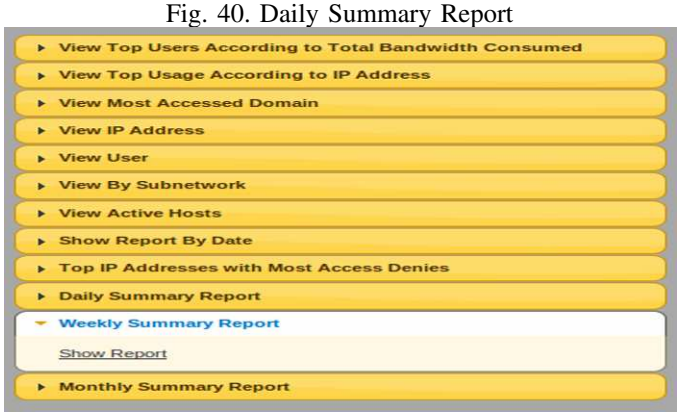
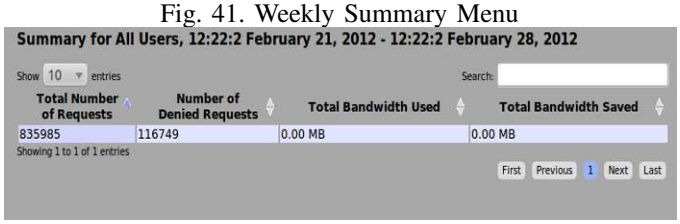


Fig. 41. Weekly Summary Menu



Summary for All Users, 12:22:2 February 21, 2012 - 12:22:2 February 28, 2012

Show 10 entries Search:

Total Number of Requests	Number of Denied Requests	Total Bandwidth Used	Total Bandwidth Saved
835985	116749	0.00 MB	0.00 MB

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Fig. 42. Weekly Summary Report

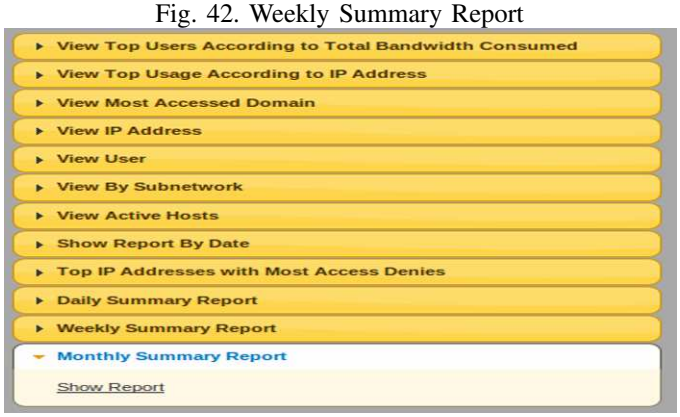


Fig. 43. Monthly Summary Menu

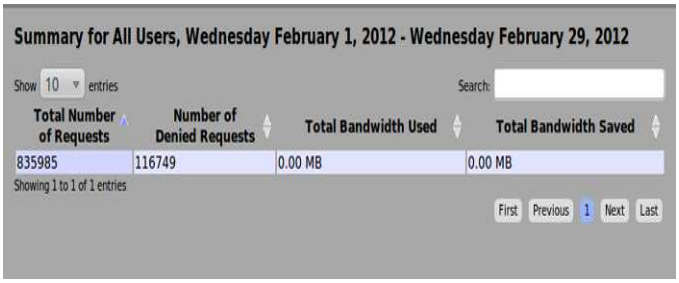


Fig. 44. Monthly Summary Report

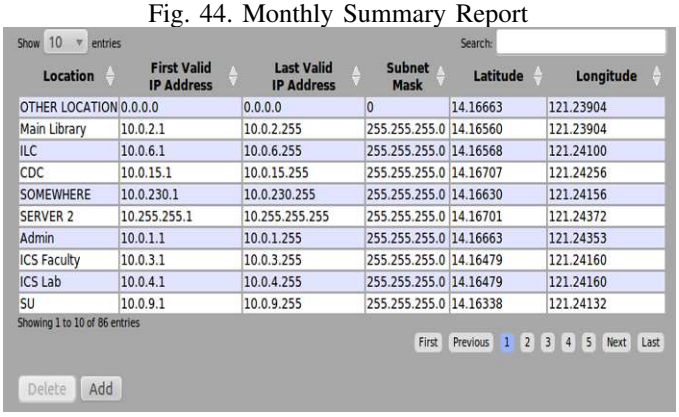


Fig. 45. Displays the currently saved list of subnetworks

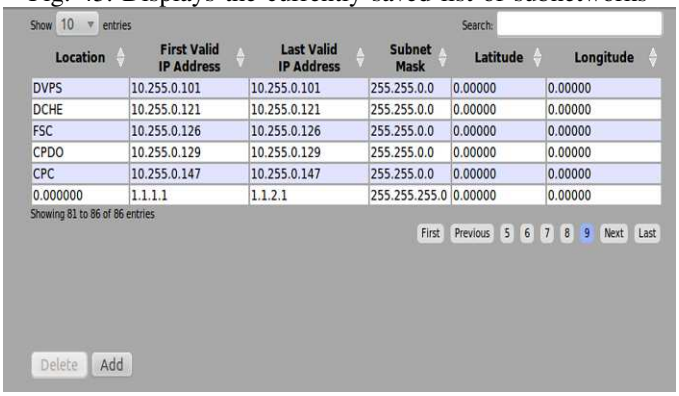


Fig. 46. Displays the currently saved list of subnetworks

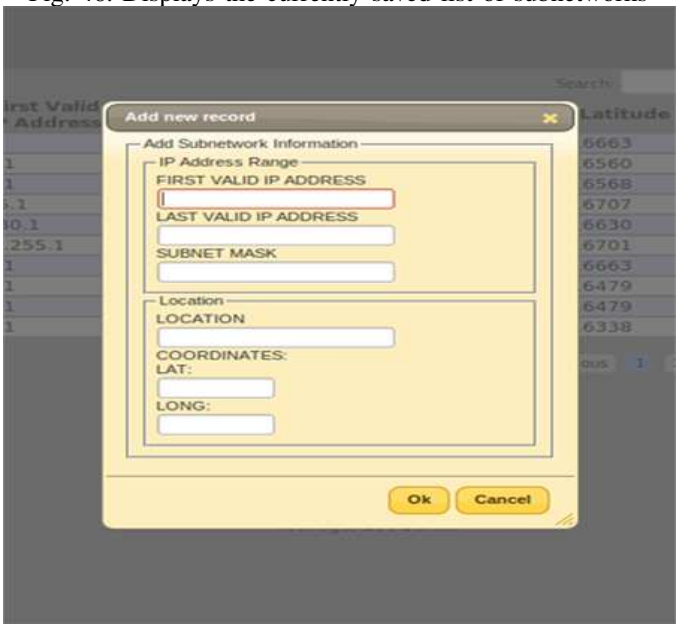


Fig. 47. Adding a new Subnetwork

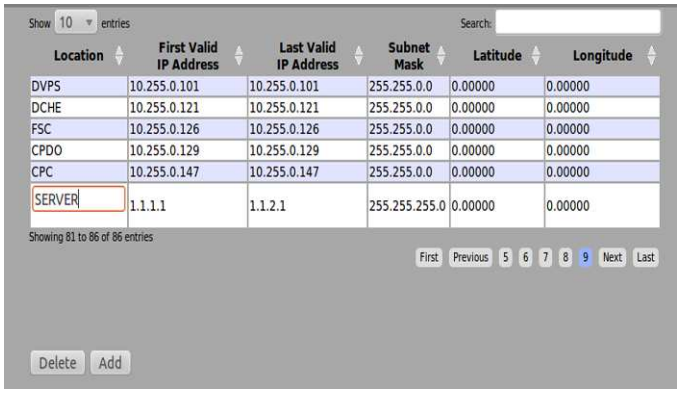


Fig. 48. Editing data about a Subnetwork

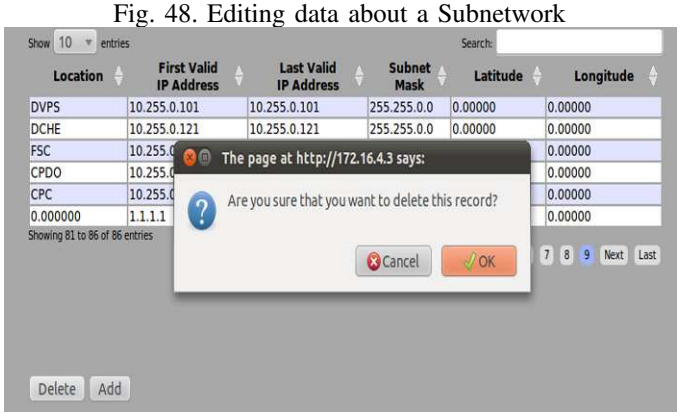


Fig. 49. Deleting a Subnetwork

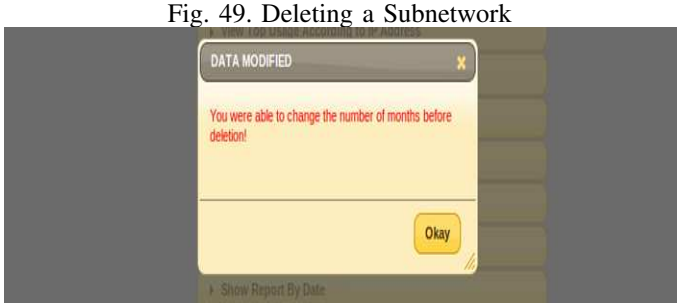


Fig. 50. Change number of Months before deletion of Logs

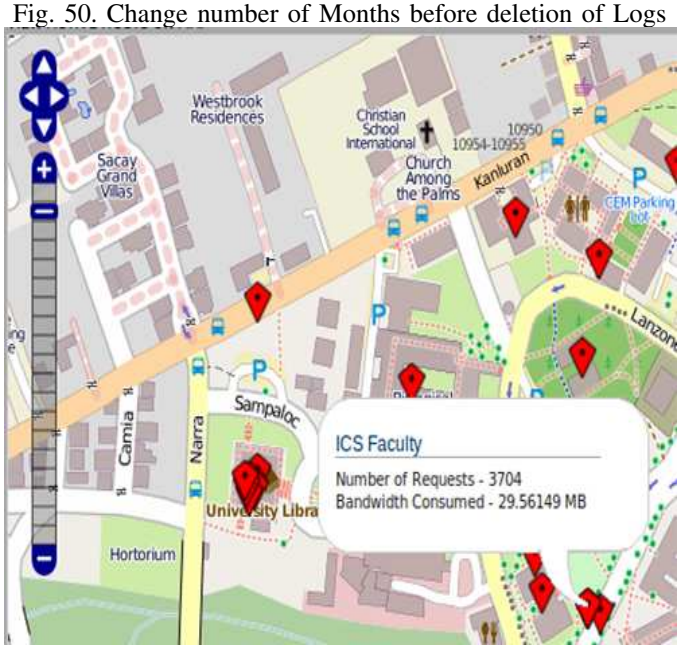


Fig. 51. Locator

VI. Conclusion and Future Work

SquiGIS was able to provide a Squid log parser using PHP 5.0 to access the Squid *access log* file. It provided a monitoring module which accesses the log files and stores the extracted information into the database. It also provided a web-based user interface for the monitoring system and added search functionalities specified by the UPLB ITC making it a more personalized Squid log parser-analyzer for the UPLB Network. It was also an effective tool to monitor the activities in the UPLB Network.

Although the system was a very helpful tool in monitoring the activities in the UPLB network, future studies may be conducted to improve on the performance of the system. A possible improvement would be to implement real time log access where the system can track each incoming request and store it to the database. Another possible improvement is to allow other squid log formats as specified in the Squid configuration file. It is also recommended to allow access to more than one server.

Acknowledgment

Shiela Kathleen L. Borja would like to thank her Adviser Prof. Joseph Anthony C. Hermocilla, Mr. Ludwid Johann B. Tirazona, ITC Systems Administration team, family and friends for the support and guidance in her Undergraduate Special Problem.

References

[1] (2011) Lightsquid home site: Home. [Online]. Available: http://lightsquid.sourceforge.net/

[2] (2011) Rapidshare ag, cham, switzerland. [Online]. Available: https://rs512l36.rapidshare.com/#!download—512l35—367652171—pdfpanda.com-install-squid.pdf—244—R 987E3E67F79BF B392155DD3945F3F44C

[3] (2011) Log analyzer tools. [Online]. Available: http://www.mela.de/Unix/log.html

[4] *Squint Installation File(install)*, Leading Edge Business Solutions (Pty) Ltd, 2011.

[5] L. J. B. Tirazona, private communication, 2011.

[6] (2011) Squid:optimising web delivery. [Online]. Available: http://www.squid-cache.org/Scripts/

[7] K. Bommepally, T. Glisa, J. Prakash, S. Singh, and H. Murthy, "Internet activity analysis through proxy log," in *Communications (NCC), 2010 National Conference on*, Chennai, India, Jan. 2010, pp. 1–5.

[8] (2011) Squidfaq/completefaq - squid web proxy wiki. [Online]. Available: http://wiki.squid-cache.org/SquidFaq/CompleteFaq#SquidFaq.2BAC8-SquidLogs.Squid_Log_Files

[9] (2011) Rotating squid logs — digital boundarygroup. [Online]. Available: http://www.digitalboundary.net/wp/?p=225

[10] G. Barnbrook. (2002) Defining a language: A local grammar of definition sentences. e-book preview. [Online]. Available: http://books.google.com/books ?id=gxPfpbVwNqsC&printsec =frontcover&dq=Defining+A+Language:+ A+local+Grammar+of+definition+ sentences&hl=en&ei=96 9oTsLNFMiGrAec35DR Cg&sa=X&oi=book_result&ct =result&resnum=1&ved=0C CkQ6AEwAA#v=one page&q&f=false

[11] R. M. John Hoftcroft and J. Ullman, *Introduction to Automata Theory, Languages, and Computation Second Edition*. 23-25 First Lok Yang Road Singapore 629733: Pearson Education Asia Pte Ltd., 2001.

[12] K. Clarke. (1995) Getting started with geographic information system second edition. e-book sample. [Online]. Available: http://www.vepraskas.com/CL341/gettingstarted.pdf

Shiela Kathleen L. Borja She is an undergraduate student under the BS Computer Science program of the Institute of Computer Science, University of the Philippines Los Baños. She is a native of Angustia, Nabua, Camarines Sur.