Academic Journal of Nawroz University (AJNU), Vol.11, No.4, 2022

https://doi.org/10.25007/ajnu.v11n4a506

Design a Web Analytics Software Tool in Order to Identify and Recognize the Behavior of Users That Visit Website

Sarhad Baez
Computer Science Department, Soran University, Kurdistan Region –Iraq

ABSTRACT

Until now we spend most of our time online using the social networking news search engines, portals and websites, we visit several websites we browse the pages and look at the information we are looking for and look for it, so it is very important to analyze it, the goal of our paper is to design and implement a software tool in order to analyze Apache log file of a website. In this paper, a software tool is created that is called web analyzer software tool in order to identify and know the behavior of users that visit a website. This information is used to improve the structure of a website.

Some facilities of our software tool are that lets a website owner to know the number of requests per a specific day, number of requests per week, and most visited pages, number of visitors per day, most used operating systems during a log file. Moreover, the website owner can investigate an IP to know if it is blacklisted or not and what an IP had done during a specific day and how much an IP was online and who are wants. Based on this valuable info a website can response to the requests appropriately. The software tool can find location of any IP address, which browsers, and Operating systems used by the user and the software tool can determine the new visitors to the website daily. Moreover, we can detect the countries of IP address and we show them graphically on a chart. By using this software, we can measure and analyze the website log files easily. The C# language is used for coding; Java script is used to find the location of the IPs.

KEYWORDS: Web analyzer Software, IP address, Web browser, Apache log file, Access Logs.

1. Introduction

Web analyzer is a software tool for analyzing websites data for the purpose of measuring and collecting visitor's activities. This analyzing will help us to improve a website and its performance and also to monitor, report websites usage, and understand user behaviors. (PETERSON, 2004) Every activity that a visitor does when it opens a website until it closes it will be recorded into a file which is called a log file. Let us to explain more about log files.

Log files are files that list the actions which have been occurred, these log files take place in the webserver and also are recorded in the database of the webserver. When a user sends a request to a webserver, the webserver also responds to the request by HTTP (Hyper Text Transfer Protocol) protocol. For instance, users may ask to download file (e.g., an image or HTTP page). However, sometimes users are malicious and they try to do a cyber-crime. In this case we should effectively manage a webserver; it is necessary to get a feedback about the activities of users. We should monitor the performance of the server and as well as any problem that may occur. The Apache HTTP server provides very comprehensive and flexible logging capabilities. Log files contain information about a User Name, IP Address, Time Stamp and Access Request, number of Bytes Transferred, Result Status, and URL that referred. Let's more talk about the contents of a log file.

The Log files record different types of information in different webservers. The basic information's present in the log file are:

- User name: This is identification of the visitor who visited the website. Mostly identification is an IP address which is assigned by Internet Service Provider (ISP).
- Visiting Path: The path taken by the users as they are visiting the web site. This may be by using the URL directly or by clicking on a link or through a search engine.
- Path Traversed: This identifies the path taken by

the user within the web site using the various links.

- Time stamp: This is a time which user spent in each web page while surfing the through the web site. Time stamp is identified as the session.
- Page last visited: The last page which visited by users before leaving the web site.
- User Agent: User Agent gives information about browser which is used by visitors to send the request to the web server. The information contains name, version and the type of browser.
- URL (Uniform Resource Locator): The resource of a web site which accessed by the user. It may be an HTML page, or a script.
- Request type: Identifies which method is used by a user to make a request (transferring information). The methods like GET, POST. (L.K. Joshila, V.Maheswari, & Dhinaharan, 2011)

A log file can be normally located in three different places:

- Web Servers
- Web proxy Servers
- Client browsers

2. System Modeling and Analysis

The system is a collection of an interrelated components that work together to achieve a purpose. System analysis is referred to the systematic examination or detailed study of a system in order to identify problems of the system, and using the information gathered in the analysis stage to recommend improvements or a solution to the system. (Kumar, 2016)

2.1 System Analysis

System analysis is the study of sets of interacting entities, including computer systems analysis. This field is closely related to requirements analysis or operations research. It is also "an explicit formal inquiry carried out to help someone identify a better course of action and make a better decision than he might otherwise have made. (Kumar, 2016)

2.2 System Modelling

During the system requirements and design activity, systems may be modelled as a set of components and relationships between these components. These are normally illustrated graphically in a system architecture model that gives the reader an overview of the system organization. System modelling helps to give more detailed system specifications which are in form of graphical representations that can describe problem to be solved or the system that is to be developed. Examples of such modelling tool is a System Flowchart.

2.3 System Flowchart

System flowchart is a type of diagram that represents an algorithm or process, showing the steps as boxes of various kinds, and their order by connecting these with arrows. This diagrammatic representation can give a step-by-step solution to a given problem. Process operations are represented in these boxes, and arrows connecting them represent flow of control. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields. Different symbols are used in the flowchart to represent input, output, decision, connectors and process. (Kumar, 2016).

3. Implementation and Results

3.1 Main Form

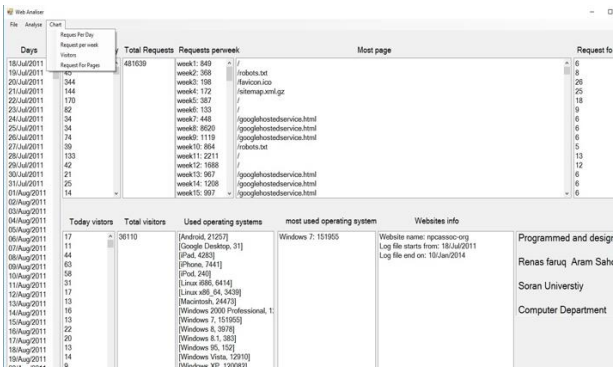


Figure 3.1 Main Form

After you open software tool you will see a Main Form as it shown with figure 3.1 Our software tool's goal is analyzing. And analyzing is divided into two parts

general and a specific part. The main form is analyzing log file data generally for that we click file-open in a menu list and open a log file to load it to the software tool.

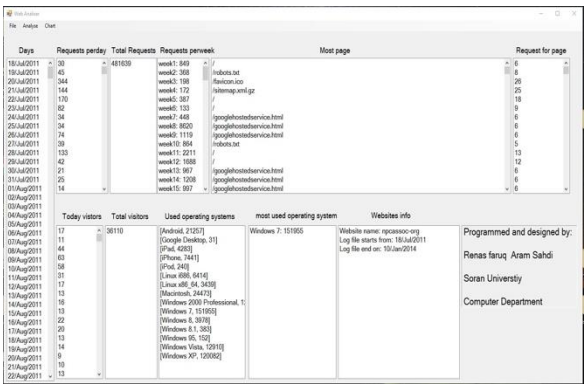


Figure 3.2 The figure shows analyzed data in list boxes

In figure 3.2 generally analyzed items will presented as bellow.

- Days: shows all days of a log file. Request per day: shows requests per day.
- Total request: shows all request from the start to end of log file. Request per week: shows all requests per weeks.
- Most Page: shows most requested page per day.
- Request for page: shows the total of requests for most page per day. Today visitors: shows the number of visitors per day.
- Total visitors: shows the total number of visitors during a log file.
- Used operating system: shows the list of operating systems which used by visitors. Most used operating system: shows the most used operating system.
- Website info: shows the name of the website and its start and end date of a log file.

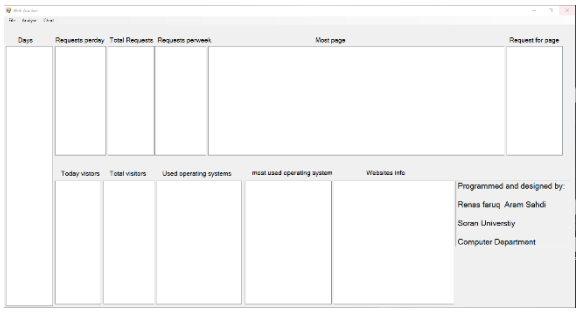


Figure 3.3 clicking on chart menu list

By clicking on chart menu list as it shown in Figure 3.3 a list will open which include types of charts:

- Request Per Days
- Request per week
- Visitors
- Request for pages

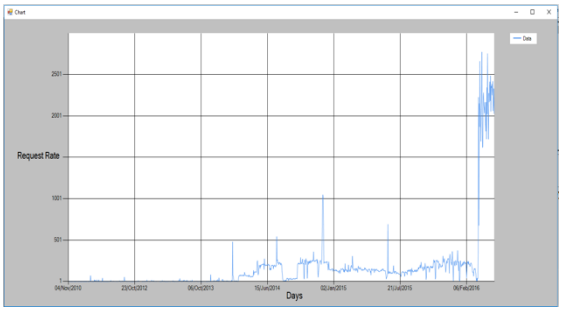


Figure 3.4 this chart show request rate per days

This chart show request rate per days as it shown in figure 3.4 of entire log file. For example, from 04/nov/2010 the requet per day rate is 1 and 15/jun/2014 is 205 and this example shows that the request at beginning of the logfile was low then after two years it started to high. The chart has ability to zoom in for looking at data with details.

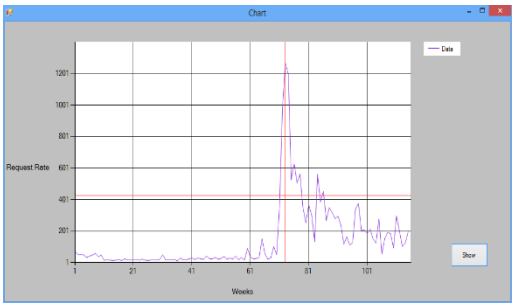
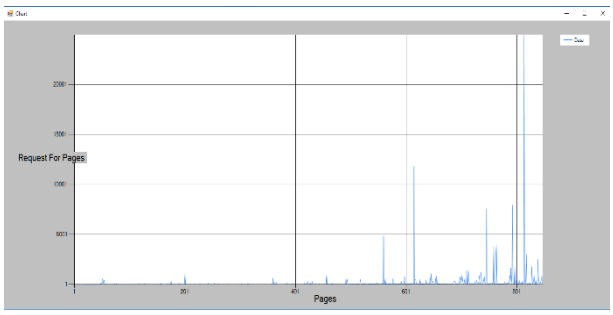


Figure 3.5 Chart per weeks

This chart show request rate for requests per weeks as it shown in figure 3.5.

Figure 3.6 chart for most pages.



This chart shows the request rate for most pages as it shown in figure 3.6.

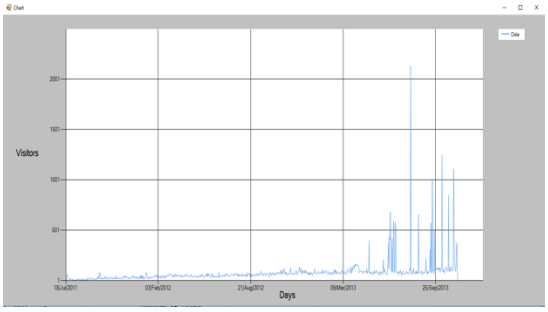


Figure 3.7 This chart show number of visitors per day

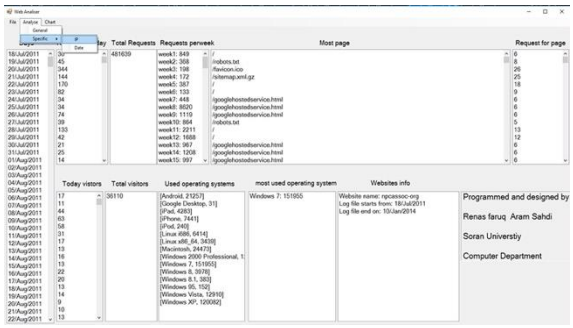


Figure 3.8 Then you get a second form which is IP part

3.2 IP Form

Then you get a second form as it shown in figure 3.8 which is IP part.

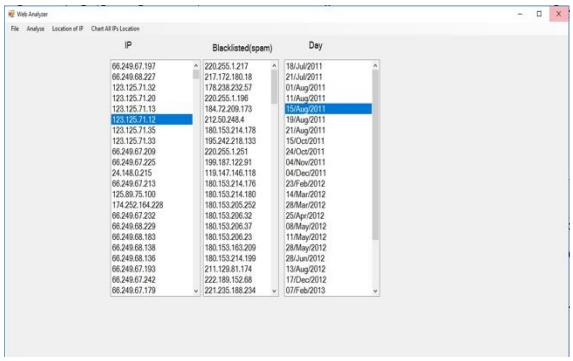


Figure 3.9 which list all unique IPs of a log file and blacklisted IPs.

This chart shows the rate of visitors per day as it shown in figure 3.7.

After analyzing generally, we can click analyze-specific on a menu list to work on logfile data specifically and specific has two items

- IP
- Date

For IP part we click on IP on menu list:

we can click file-open to load log file to a listbox as it shown in figure 3.9 which lists all unique IPs of a log file and blacklisted IPs.

IP: IP listbox contains all unique IPs of entire log file.

Blacklisted IPs(spam): shows those IPs which identified as blacklist(spam) by https://myip.ms/browse/blacklist and the reason for blacklisting is they might have done something against law like ddos attack or any cyber attacks.

Day: when we click on any IP it includes all days which selected IP have been connected to the website.

Then for analyzing we select one IP and one day which have connected to the website and click on Analyze so the analyzed data will be the activity of a user for only selected day.

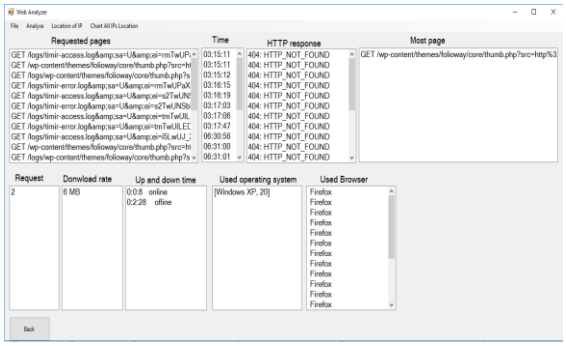


Figure 3.10 then for analyzing we select one IP and one day

After selecting a specific IP with a date and clicking on analyze as it shown in figure 4.12 analyzed data will shown

- Requested pages: Shows all pages that have been requested by the specified IP and the example shows that IP requested 20 pages
- Time: Shows exact time of each page that have been requested
- HTTP response: contains web server response for each requested

  page and servers response to the requests of that IP were (404: HTTP_NOT_FOUND) in that example
- Most page: shows most opened page that requested by the specified user
- Request: shows how many times most page requested that was (2) in that example

- Download rate: includes the amount of data that was downloaded by the user which was (6Mb) in that example
- UP Time: The uptime represents a time that a user starts communication with the server until he terminates the communication with the server, which have been (8 seconds) in that example
- Downtime: The downtime represents the interval time from the time that a specific user disconnects from the server until the time he connects again to the server. Which was (2 minutes and 28 seconds) in that example
- Used operating system: Represent the operating system that used by the selected IP during the selected day. Which is (windows XP)
- Used Browser: Represent the browsers used for the selected IP and mostly browser used is (firefox) in this example.

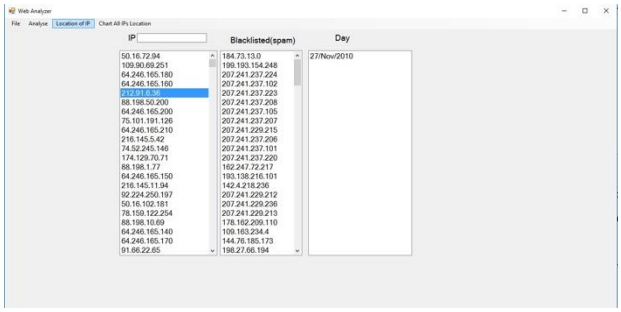


Figure 3.11 location of IP address

Select a specific IP Address then click on Location of IP as it shown in figure 3.11.

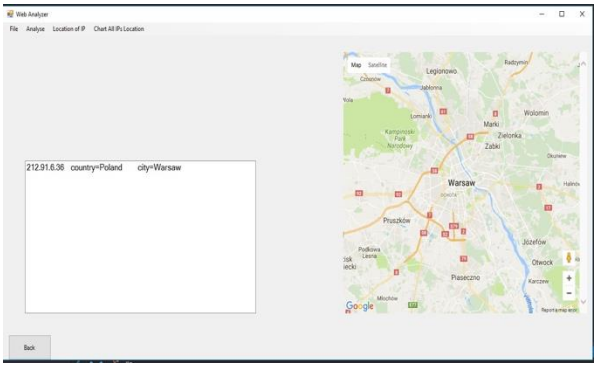


Figure 3.12 showing geolocation of an IP on the map

A text filed Shows the IP Address and the Country that belongs and the City as it shown in figure 3.12 and There is a default C# web browser tool which loads a google API that shows the City location of the IP address.

The example shows this IP 212.91.6.36 which is from Poland and Warsaw city.

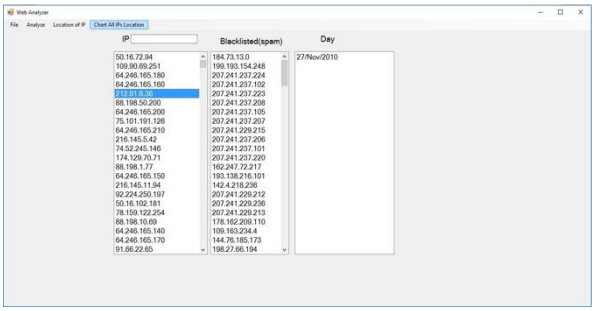


Figure 3.13 All IPs Location chart

By clicking on Chart All IPs Location as it shown in figure 3.13 it loads all unique IPs to the new form

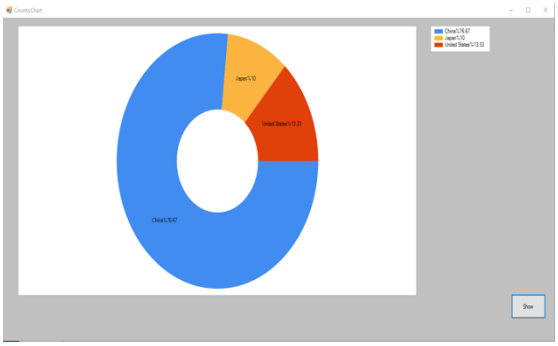


Figure 3.14 Chart for all country of IPs

By clicking show button in this form will this chart will show us Country of all IPs as shown in figure 3.14 and show us the percentage of each country that unique IPs belong to. In this figure we got a chart which shows that 76% IPs where from China, 13% from USA and 10% from Japan.

3.3 Date Form

In the Main form you can click on analyze-specific-Date and you will get a new form and you can open and load log file data to software tool and you will get all days in a list box.

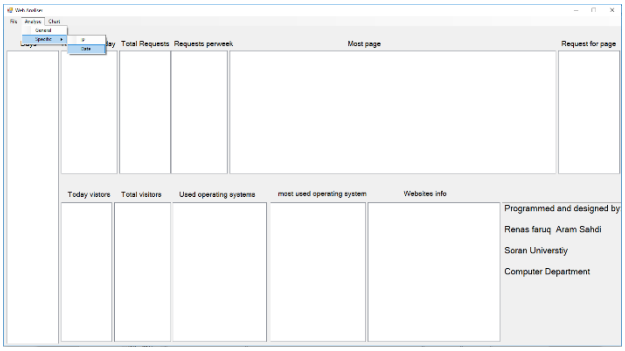


Figure 3.15 clicking on Date

By clicking on date, the new form will be opened as shown in figure 3.15.

Figure 3.16 analyzing date data

After opening a file as it shown in figure 3.16 there is a list box which contains all date of entire log.

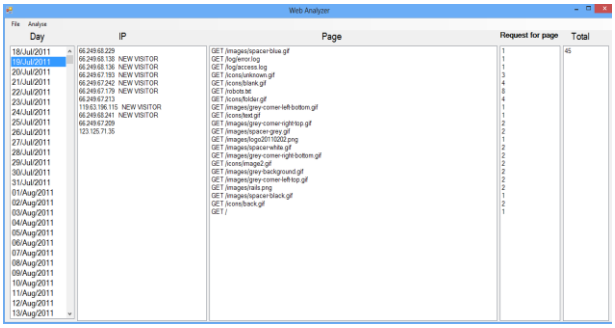


Figure 3.17 analyzing specific day

After that you can select one day as it shown in figure 3.17 and click on analyze to know what has happened on that day.

- Day: shows all days of a logfile and in this example, we selected 19/Jul/2019.
- IP: shows all unique IPs which connected to the website and if they are new visitor or not and this example shows that 11 visitors connected to the website and 7 of them are new visitors.
- Page: shows all pages requested by the visitors during that selected day.
- Request for page: shows the number of requests for requested pages in this example we see that page: robots.txt requested 8 times.
- Total: shows the total of requests and in this example, we see that we have 46 requests during that selected day**.**

4. Conclusions

The web analyzer software tool is designed to improve the structure of a website and the security of the website. It helps the website owner, admin and developer to have a better point of view of their website through collecting data for a month or a year. These data are analyzed through our tool.

Understanding log file of the website is hard and it needs time to look at it; therefore, our tool helps to reduce time, helps understand easily and get results to user. The facilities such as all days of log file, requests per days, request per weeks and most requested page of a day, and visitors per day, with showing operating systems and browsers used to open the website, make the opportunity for the website's owner to understand it easily. There by the owner of website will have a clean view of the goal for achieving improvement of website by knowing what users do and what their favorite is.

5. References

1. Shakti Kundu ; L Garg (2017). Web log analyzer tools: A comparative study to analyze user behavior, 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence
2. AVINASHKAUS. (2009). Web Analytics 2.0. Wiley Publishing,Inc.
3. Catherine, D., Yi, Z., & Starr, R. (2004). Using Web Analytics to Measure the Activity in a. Dwyer et al., 11. Clifto, B. (2009). Advanced Web Metrics with Google. John Wiley & Sons, Inc.
4. Farhana S, Hania A, Hamid M, Sana K (2019). Browsing Behaviour Analysis using Data Mining, International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2,
5. H. (2006). Measuring the Success of Your Website: A Customer Centric Approach to Website Measurement. Person Education Australia.
6. Jansen, B. J. (2006). Search log analysis: What it is, what's been done. Library & Information Science Research, 26.
7. Kumar, P. D. (2016). Overview of System Anaylsis and

Design.

8. L.K. Joshila, G., V.Maheswari, & Dhinaharan, N. (2011). ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING. International Journal of Network Security & Its Applications (IJNSA), 12.
9. PETERSON, E. T. (2004). WEB ANALYTICS DEMYDFIED.
10. Sterne, J. (2002). Web Metrics: Proven Methods for Measuring Website Success. New York: John Wiley & Sons,Inc.
11. Strickland, J. (2011). Systems Engineering Processes and Practice. Lulu press.
12. Suneetha, K. R., & Krishnamoorthi, D. R. (2009). Identifying User Behavior by Analyzing Web Server. IJCSNS International Journal of Computer Science and Network Security, 6.

1.