Al-Imam Muhammad Ibn Saud University

College of Computer and Information Sciences

Information Systems Department

Threats to Web Data Analytics and how to Prevent them

By

Sara Mnour Almutairi

Thesis submitted in fulfillment of the requirements
For the degree of
MSc

April 2015

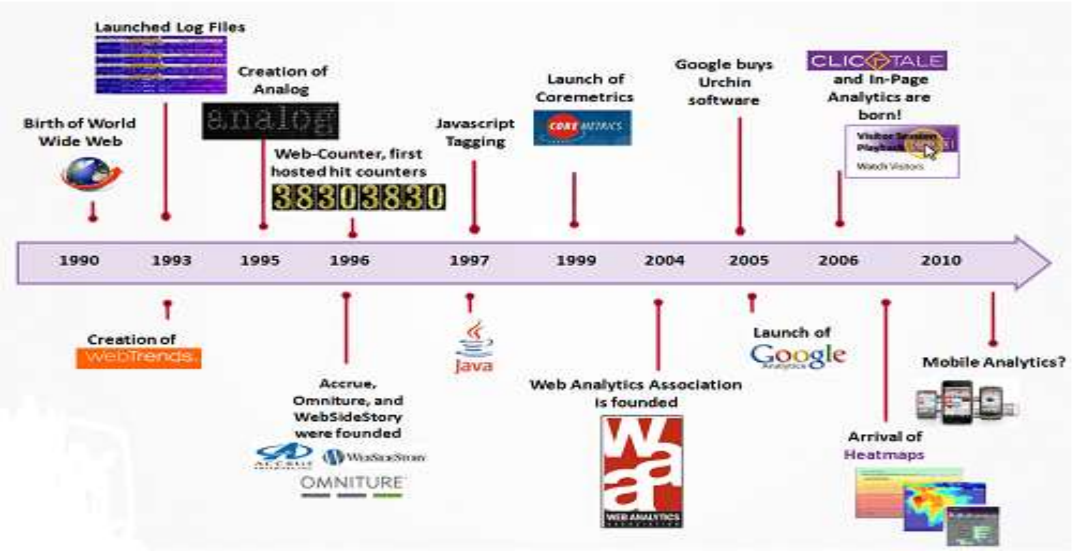
Introduction

Web analytics is a way to maintain and track the traffic and the user's acts on websites through specific algorithms and procedures. It supports and enhances the collaboration between experts and users through the use of an online application. Web analysis is an ever existing need for businesses, organizations, and individuals regardless of their domain or field of work. Since businesses are continuously growing, they tend to adapt new strategies to develop.

Definition

Einstein said – "Not everything that can be counted counts and not everything that counts can be counted." Web analytics is the search for what counts. Web analytics are tools and methodologies used to enable organizations or users to track the number of people who view their site per period of time, such as how many visits of unique users, how many total visits, or it can be per subject of the web, such as how many users visited a given page, clicked a specific link, or viewed a given subject, and then use such information to measure the success of their online strategy and to assess and improve the effectiveness of a website

There are many tools to analyze the web traffic; one common tool is the Google Analytics, which we will thoroughly use in our demonstration.



Usage and Importance

As time emerges, web analytics are becoming more essential especially for businesses, as they can provide an insight into the performance of a website. Many Web hosting packages include Web analytics tools. Free services such as Google Analytics, which can be used in any site, are also popular. Companies use Web analytics to find out information about their visitors, including how they interact with the pages in a site. The information gleaned from analytics can help to inform future decisions regarding the content and marketing of Internet and other company services.

Problem Statement: Threats to Web Data Analytics

As noted, web analysis became very essential in many aspects, and for now there are some great tools that use different methodologies to achieve this goal. However, like any method, there are some weaknesses for each methodology, and the common one is that none gives one hundred percent accuracy, which may become a severe problem, and affects the decision the user may take if the problem was not recognized. Besides that, the analytical tool depends on some components on the PC such as using the JavaScript or using IP addresses. These would lead to inaccurate data analysis if the JavaScript is disabled, or if one IP address is shared by many users, or even if several IP address are used by one user. These issues are the topic of this thesis. These are the threats to Data Web Analytics, and this topic will discuss how to mitigate and prevent them.

Research Questions

The goal is to increase the accuracy of the analytics and prevent the threats, so the main research question is:

- How can threats to web data analytics be prevented so as to increase the accuracy?

  Other minor questions that the thesis will answer are:

- What are the issues that prevent such accuracy?
- How to overcome these issues, or minimize their effects?

- Is there a way to get full accuracy and prevent such threats, theoretically or technically?

Research Objective

The main objectives of this research are as follows:

1. Suggest solutions, theoretical or practical, that can be used to spotlight some existing ways that are not utilized, or by giving new ideas.
2. Enumerate some of the current threats, and the solutions that are currently used to overcome them.
3. Discover if there is a methodology that may produce one hundred percent web analytics accuracy.

Scope of the thesis

This thesis intends to take a look at the web data analytics from all aspects. It starts with a definition of the analytics, and moves to the history. Also will talk about how it was developed, the current state and tools used, what are the main usage of the analytics, the advantages and disadvantages of the current tools and methodologies, and concluding with some ideas and theories to enhance the current methodologies to raise the accuracy of the analytics done.

Significance of the Study

Data analytics is general is used in wide areas, in financial world, military world, governmental work, all depends in the data they collect, and their decisions are based on the data in their hands.

Data web analytics is vital for websites searching for more traffic, more income, or for launching marketing campaigns, and much more aspects. Most of the studies conducted on threats to web analytics had focused on giving solutions to particular threats, what we will do in this study, is to try to give one solutions that is applicable for more one threat at one time, this will help both programmers and analysts by reducing their work to use different logging methodologies, and comparing data from more than one analytical resources respectively.

Thesis Organization

The thesis is organized into chapters, each with a unique general title. Each chapter is divided into sub titles, each with a title that summarizes its main idea. The sub-part may be divided into other sub-parts with titles and so on.

Research Design

In order to complete our experiment, we have divided our work into the following steps:

1. Collecting and analyzing threats
2. Testing analysis with some current threats
3. Suggesting new solutions
4. Testing the new solutions

Each step may be divided further into sub-steps that will be mentioned separately when encountered.

Theoretical background

Key Words in Web Analytics (Web Analytics Metrics)

In order to understand the benefits of Website analysis, one must first understand metrics – the different kinds of available user information. Although the metrics may seem basic, once collected, they can be used to analyze Web traffic and improve a Website to better meet its traffic. According to (Panalysis n.d.) an Australian Web analytics company, these metrics generally fall into one of four categories: site usage, referrers (or how visitors arrived at your site), site content analysis, and quality assurance. Although the type and overall number of metrics vary with different analytics vendors, there is still a common set of basic metrics common to most. Table 2 outlines eight widespread types of information that measure who is visiting a Website and what they do during their visits, relating each of these metrics to specific categories. There are no globally agreed definitions within web analytics as the industry bodies have been trying to agree on definitions that are useful and definitive for some time. The main bodies who have had input in this area have been JICWEBS (The Joint Industry Committee for Web Standards in the UK and Ireland), ABCe (Audit Bureau of Circulations electronic, UK and Europe), The DAA (Digital Analytics Association), formally known as the WAA (Web Analytics Association, US) and to a lesser extent the IAB (Interactive Advertising Bureau). However, many terms are used in consistent ways from one major analytics tool to another, so the following list, based on those conventions, can be a useful starting point. Both the WAA and the ABCe provide more definitive lists for those who are declaring their statistics as using the metrics defined by either.

Threats to Data Web Analytics

As noted so far, the main objective of analyzing web traffic is to figure out the visitors' behaviors, in order to set future goals for business. These goals and strategies depend heavily on what we'd get from the analysis. Thus, such analysis must be accurate, or at least if not hundred percent accurate, should determine the weaknesses to act accordingly.

Despite the pitfalls, error bars remain relatively constant on a weekly, or even a monthly, basis. Even comparing year-by-year behavior can be safe as long as there are no dramatic changes in technology or end-user behavior. As long as the same yardstick is used, visitor number trends will be accurate. For example, Web analytics data may reveal patterns like the following:

- Thirty percent of site traffic came from search engines.
- Fifteen percent of site revenue was generated by product page x.html.
- Subscription conversions increased from our email campaigns by 20 percent within a week.
- Bounce rate decreased 10 percent for our category pages within one month.

With these types of metrics, marketers and webmasters can determine the direct impact of specific marketing campaigns. The level of detail is critical. For example, it can be determined if an increase in pay-per-click advertising spending—for a set of keywords on a single search engine—increased the return on investment during that time period. As long as inaccuracies are minimized, Web analytics tools are effective for measuring visitor traffic to the online business.

Web analytics results are never one hundred percent accurate! This is the truth, and there are many factors that affects the results, for example disabled JavaScript.

This study will have a look over a list of known factors and issues that affects the collection of the data in Web analytics.

Definition of threats to data web analytics

In a nutshell threats to data Web analytics are a group of elements, whether technical or non-technical, that prevents from having an accurate measure of the Web analytics.

Collecting and Analyzing Threats

Analysis of web data face some difficulties, these difficulties are the threats, and in order to achieve solutions, we first should collect these threats, and analyze them in order to understand their causes, impacts, so it is more applicable to find solutions for them.

For this step to be done, many resources were conducted to find the threats, the resources included websites, books, white papers, previous studies, journals and more. This step was the easiest one, we have found many references that shows what the threats are, and some of the threats were:

- Dynamic IP addresses
- Disabled JavaScript
- Counting Robots
- Inability to detect unique Users

These threats will be explained more in the next chapter. When threats analysis steps reached, we found that there is no one source for the threats, as each threat may be caused by unique reason (for example counting robots reason is not the same as disabling JS, and the hardness of detecting unique reasons comes from dynamic IP addresses, or that the user may visit the website from new location). Moreover, some threats can be divided into two categories, technical and human actions, where the technical includes for example counting robots, and human-actions includes disabling JS.

Suggesting new Solutions

This step along with the next step are considered to be the most important steps, as the solutions will be generated here (theoretically) and then in the next step and will be tested. The focus was on finding solutions that are simple but yet powerful and with few limitations as possible as can, also we set focus on giving at least one solution that can overcome many threats at once, keeping in the other hand its simplicity, reliability, and the ability to implement it with little modifications to the current technology.

The solutions that were generated are:

- Amending the URL with extra information
- Using the HTTP_USER_AGENT
- Using Social Media
- Login to browser

Issues in this step: Many solutions came in our hand, but after analyzing them we found that either they have many limitations, or that they may violate terms, such as privacy terms at a high level, for example on solution was to try tom combine the IP address along with the location of the user. Such solutions were filtered and eliminated.

Testing the New Solutions: How to Prevent Threats to Data Accuracy

Since the problems affecting the accuracy are not related to each other, everyone will be countered alone and suggestions and solutions for it will be proposed. In case the solution is applicable for more than one problem, it will be stated.

Once the data are logged, the analytical tools will analyze the data as is. If the data sent is wrong, that is not accurate, then the analysis result will be inaccurate, and vice versa. So, we will emphasize in generating solutions that will deal with the data sent from the client side.

If a solution needs coding, we will use the PHP scripting language for the server side, and JavaScript for the client side.

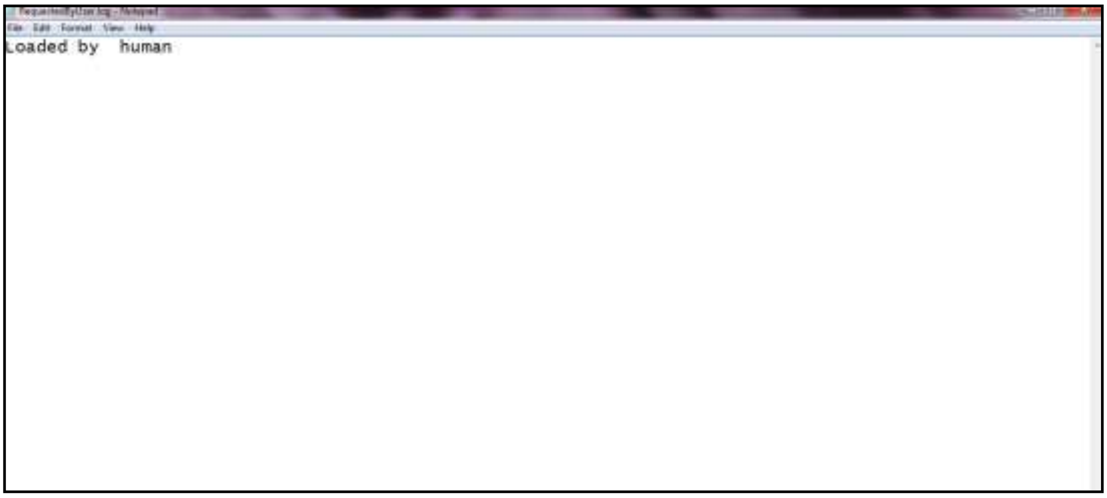
Adding RequestByUser

Experiment

The experiment was conducted on a simple PHP web page created, with a code to check for the existence of the variable RequestByUser. For this experiment, a simple C# browser was created, and a crawler was used.

Results

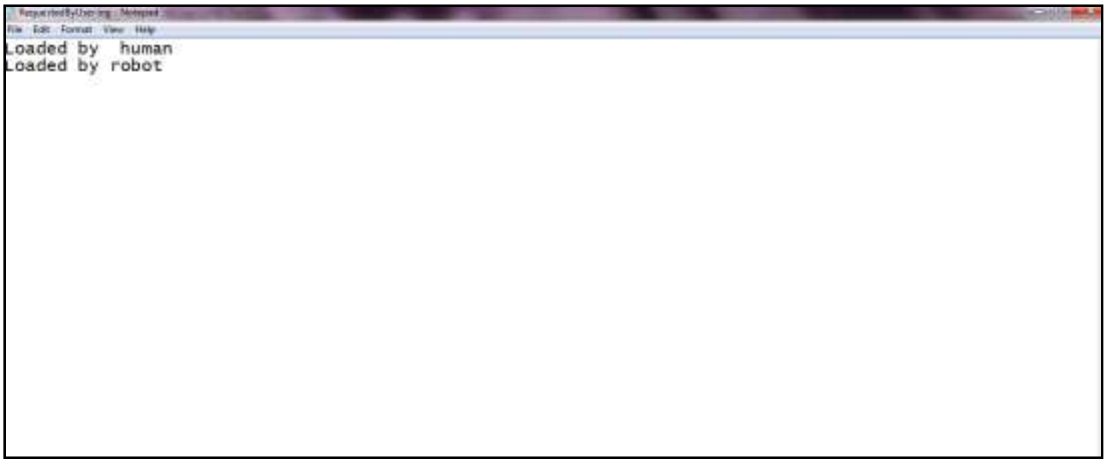
After the URL is typed in the navigation bar and clicked navigate, the page sent logging data to the RequestByUser.log file, the files contains



```
Loaded by  human
```

Figure 5.1 RequuestByUser.log -1

The second step was to request the same page from the crawler, after the crawler finishes its work, the RequestByUser.log file will contains:



```
Loaded by  human
Loaded by robot
```

Figure 5.2 RequuestByUser.log -2

The page logged that it was requested by a robot, the reason behind this is that the variable RequestByUser was not set as was done in the first step.

RequestedByUser variable will allow to check whether the URL is called from a link that was clicked by a user, or typed directly into the browser's URL textbox, or if it was requested by a robot, this is applicable because the browser will add this variable by itself, this as mentioned will allow to distinguish real users from robots, so the accuracy of the analytics will increase, and the data provided to the analyst can be separated into real visits and robots visits. Such separation will have a great impact on the decisions that are taken based on the analytics, as many studies showed that between 51% to 60% of the traffic on the web are generated by robots, this means that a chance that each two visits to any website, 1 of them might be a robot, such chance will highly trick the analysts, and a wrong decision may be taken.

Moreover, it will also allow us to overcome the issue of disabling JavaScript, since this option will not depend on JavaScript and that the user cannot change POST variables or disable them, this variable will always be sent.

HTTP_USER_AGENT

Experiment

HTTP_USER_AGENT is one of the 19 CGI environment variables defined by The National Center for Supercomputing Applications (NCSA). It is used to identify the user agent that the client is using when a CGI script is accessed from a web page.

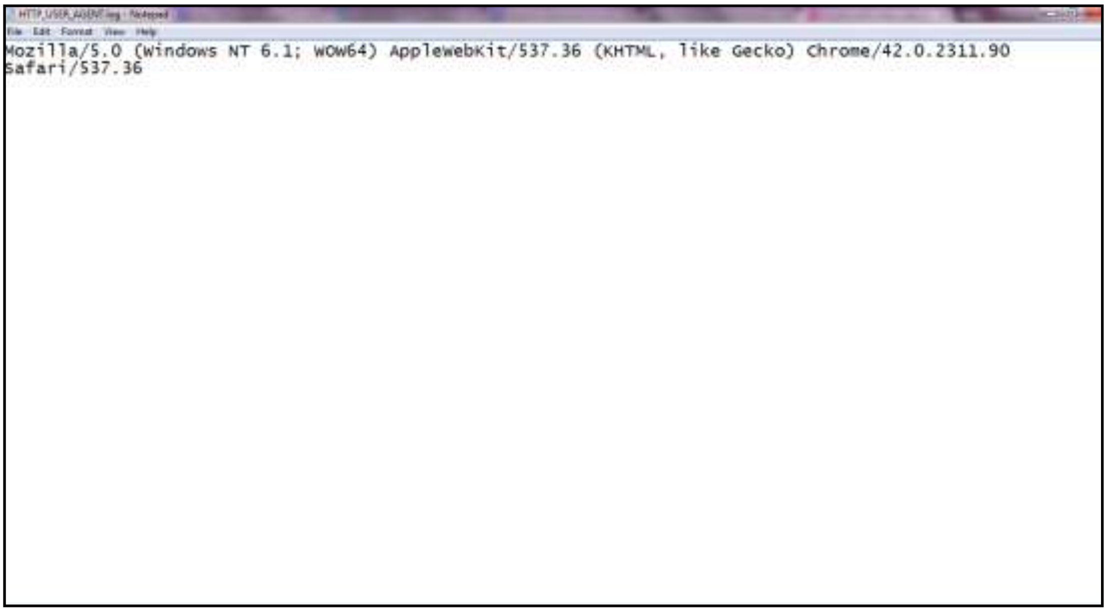
A page was created that reads this variable, and send logging data accordingly.

Results

The first request to the page done using the browser, and the second request was done using the crawler.

After the page loaded using the browser, the corresponding log file, http_user_agent.log showed the following logging data

The page was able to tell us who had requested het by just displaying to us the data it received from the HTTP_USER_AGENT variable.



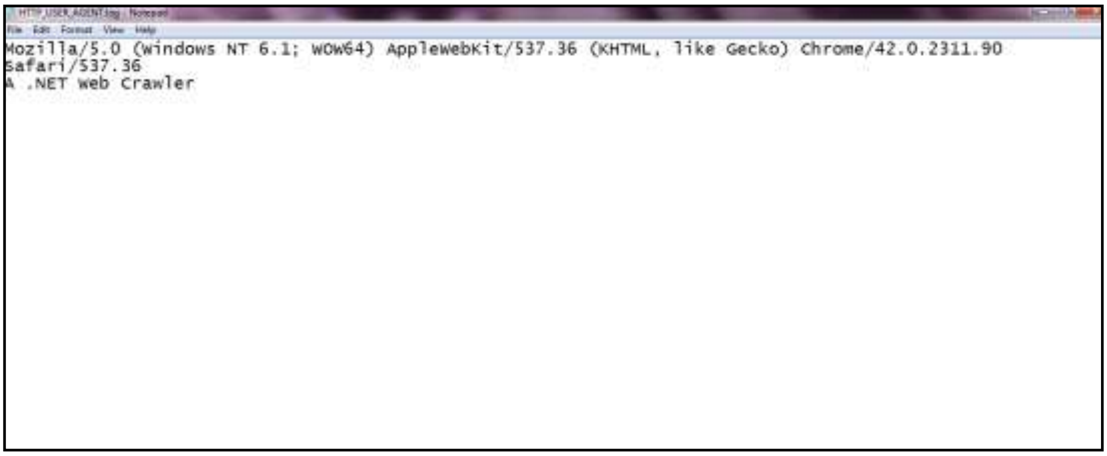
```
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/42.0.2311.90
Safari/537.36
```

Figure 5.3 HTTP_USER_AGENT.log -1

In the first line, last word before the numbers, is "Chrome" which is the browser used in requesting this page. The first word is Mozilla, not to be confused with Mozilla Firefox, is the core engine of some browsers, including Chrome, and Firefox.

From this line, it is clear that the page was requested by some program, this program send clarification data about itself, telling I am a browser.

In the second step, after the crawler requested the page, the corresponding log file shows:



```
Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/42.0.2311.90
Safari/537.36
A .NET Web Crawler
```

Figure 5.4 HTTP_USER_AGENT.log -2

It very clear how the page was able to identify the caller. When the crawler requested it, the page logged "A .NET Web Crawler”, which is the identifier of the crawler used.

The main benefit that can be acquired when using this method, is that when combined with another algorithm to filter the data, for example using regular expression to find if the user gent variable contains a name or identifier of a crawler, then this will give a clear idea, if the visit is from human, or robot

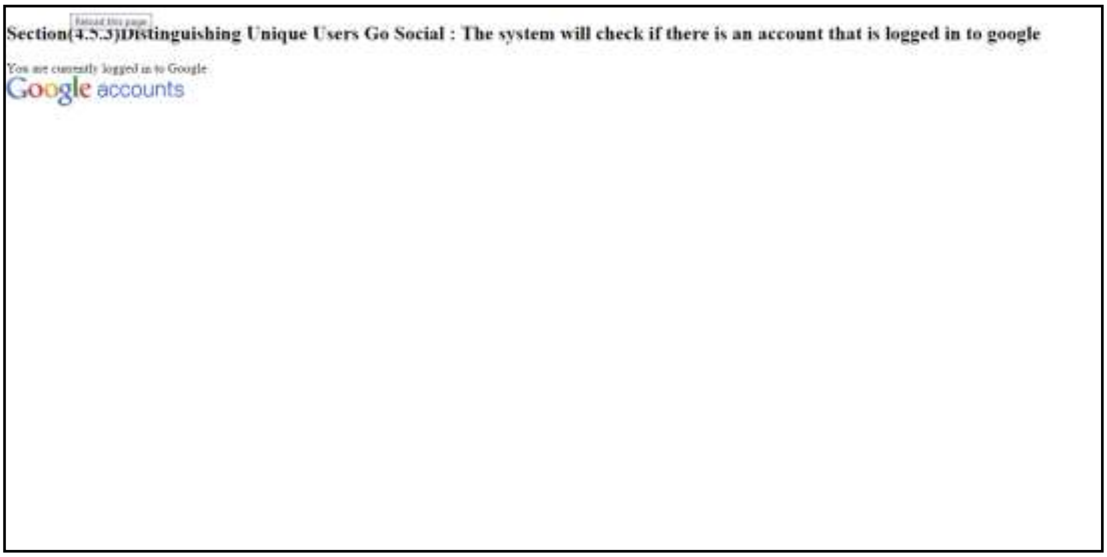
Using Social Media

Experiment

This experiment was conducted on Google, a page was created that contains a JS code to check if there is any account that is logged to Google while the pages is visited, this experiment objective is to check if there is a way that we can track and log data, using social media, which a high percentage of people use the internet while they are at the same time using social media, or emails.

Results

First we signed in to Google, using personal account, keeping signed in, the page CheckIfUserLoggedIn.html that contains the JS code was called, the result was:



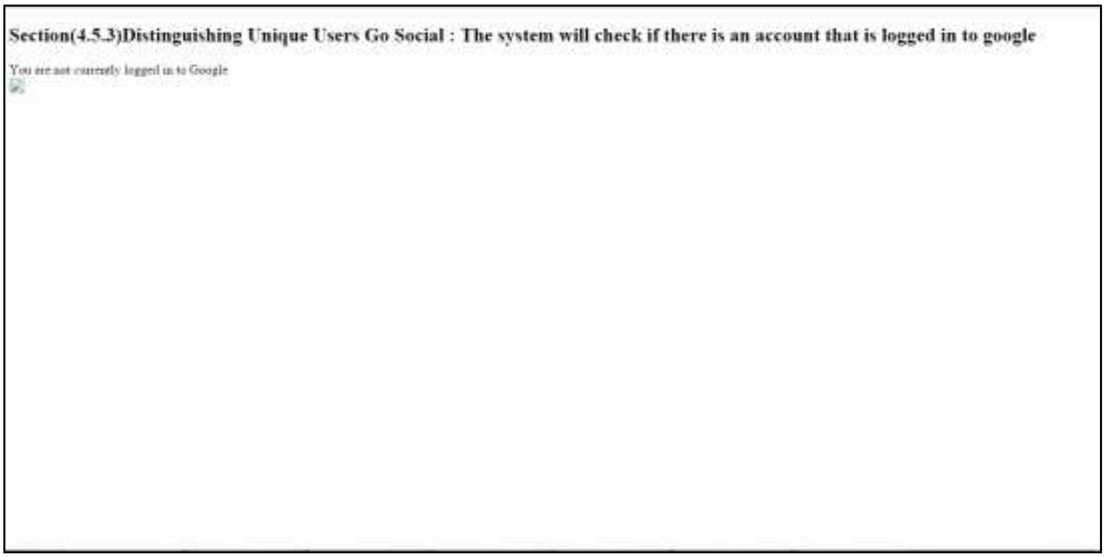
Section(4.5.3)Distinguishing Unique Users Go Social : The system will check if there is an account that is logged in to google

You are currently logged in to Google

Figure 5.4 Login with Google -1

Clear enough, the page was able to detect that there is an account that is logged in, however as a limitation, that we will suggest a solution for, the logged in email cannot be detected.

After the logging out, the same page was recalled, and the result was:



Section(4.5.3)Distinguishing Unique Users Go Social : The system will check if there is an account that is logged in to google

You are not currently logged in to Google

Figure 5.5 Logging With Google -2

Again, the page was able to detect that no one is logged in to Google.

One missing part of the result, as mentioned, is the inability to detect who is logged in. To get information of the current user, Google API should be used, however the page must include in it a request to the user to log in which is not reliable, as the user may not log in.

As a solution to this limitation, we suggest that such websites, Google, Facebook, and Twitter and so on, provide a mechanism for the developers to detect the information of the user that is logged in, without the interruption of asking the user to log in.

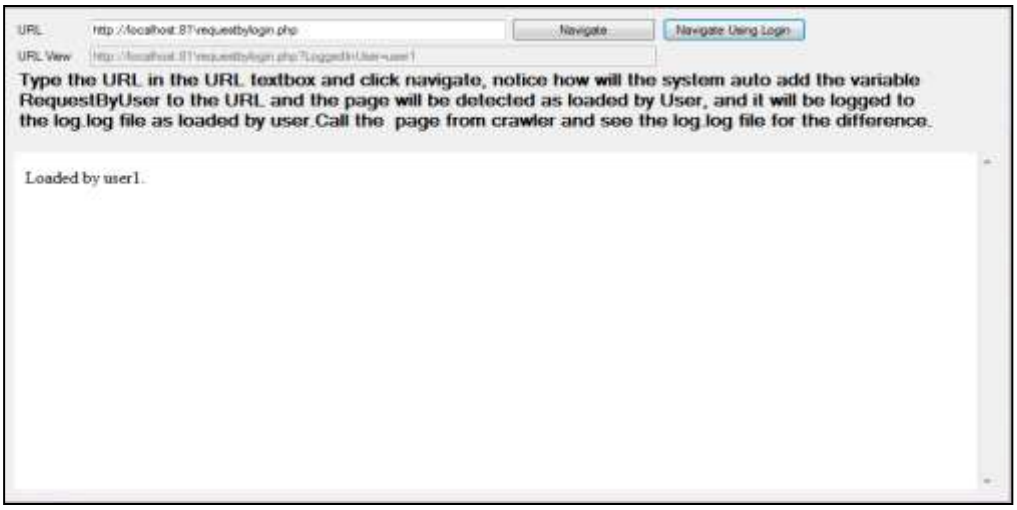
Login to the browser

Experiment

In order to detect if same user is visiting the webpage from many places or from different computers, an experiment was conducted that asks the user to login before he can use the browser, the login information is saved in the browser related storage, such as a file, or a database, in this example these information was stored in a text file.

Results

After the browser was launched, the login was made as a user that is named "user1", then a page was requested, that checks for a variable that store the username of the user, the page returned:



URL: http://localhost:81/requestbylogin.php

URL View: http://localhost:81/requestbylogin.php?LoggedInUser=user1

Type the URL in the URL textbox and click navigate, notice how will the system auto add the variable RequestByUser to the URL and the page will be detected as loaded by User, and it will be logged to the log.log file as loaded by user.Call the page from crawler and see the log.log file for the difference.

Loaded by user1.

Figure 5.6 Mini Browser -2

The page was able to say by whom it was requested, in this case by user1.

Literature Review

Since the evolution of the Web Analytics, the issue of how to increase the accuracy of the Web Data Analytics has been a huge concern. And the issue continues to grow as the online world keeps attracting more individuals and more businesses. The studies and whitepapers were aimed to spot light on the accuracy of the Web Data Analytics, how to increase it and overcome the threats affecting it.

The researchers' work depended on their perspective, and which methods they preferred to use in analysis. Most of them have not addressed all the problems against a high rate of accuracy. In fact, they gave solutions to some problems in specific, or recommendations on how to treat and use the data logged. Some others talked about when to analyze data and when not to, or gave recommendations on how to use some specific analytical tool features in order to increase the accuracy.

Web Data Analytics Studies Review

A white paper under the title of "Advanced Web Metrics Whitepaper, Understanding Web Data Accuracy" by (Brian Clifton, 2010) had revised some of the threats affecting the data analysis, such as dynamic IP addresses. After the author had explained some aspects of data accuracy he suggested some recommendations to enhance accuracy and decrease the threats, he advised the use of visit metrics in preference to unique visitor metrics because the latter are highly inaccurate.

In the section titled "Ten Recommendations for Enhancing Accuracy" he mentioned some ideas to enhance the accuracy, such as selecting a tool that uses first-party cookies for data collection.

In web analytics, each visitor has many identifiers (can be identified by cookie, IP etc...) these identifiers should not be confused, for example, if first-party cookies are deleted, do not resort to using IP address information. It is simply better to ignore that visitor. Also, Remove or report separately all non-human activity from the data reports, such as robots and server-performance monitors. To increase the accuracy, everything should be tracked. Tracking must not be limited to landing pages. The entire website’s activity, including file downloads, internal search terms, and outbound links, should be tracked (Brian Clifton, 2010).

Some analysts, used to analyze the web data on daily bases, although this may be important for some websites, however these duration analytics is mostly inaccurate.

In addition to the previous stated paragraphs, there are some recommendations (Brian Clifton, 2010):

1. Regularly audit your website for page tag completeness (at least monthly for large websites). Sometimes site content changes result in tags being corrupted, deleted, or simply forgotten.
2. Display a clear and easy-to-read privacy policy (required by law in the European Union). This establishes trust with your visitors because they better understand how they’re being tracked and are less likely to delete cookies.

3. Avoid making judgments on data that is less than 24 hours old, because it’s often the most inaccurate.
4. Test redirection URLs to guarantee that they maintain tracking parameters.
5. Ensure that all paid online campaigns use tracking URLs to differentiate from non-paid sources.
6. Use visit metrics in preference to unique visitor metrics because the latter are highly inaccurate.

The top five common threats to data web analytics (5 Threats to Data Accuracy, 2014) are:

- Robots
- Extensions blocking analytics tools
- Cookies
- Cross-device browsing
- Disabled JavaScript

In order to filter the requests that are human, among that from robots,a regular expression should be used, an example is (^(microsoft corp(oration)?|inktomi corporation|yahoo! inc\.|google inc\.|stumbleupon inc\.)$|gomez) (5 Threats to Data Accuracy, 2014).

One reliable solution would be to send analytics data yourself, on behalf of your visitors. This is called “server-side” analytics. Server-side analytics produces reliable data, but has two major drawbacks: first, the implementation can be long and painful, and will definitely require involvement from your IT team, and second you will lose valuable client data, like time spent on a page or other user interactions that don’t generate a server call in order to overcome extensions blocking analytics tools". On the other hand, the document said that regarding the disabled JavaScript, one has no control over it, although the amount of people who disable JavaScript is decreasing.

For the cross-device browser, a good solution was mentioned, use Google account, and this solution can be extended far more. Despite the wealth of useful information available in log files, the data also suffer from limitations, creating challenges for the people using them. The limitations of Web log files generally arise because certain types of visitor data are not logged, such as information about the person visiting the site rather than just the computer visiting the site, and some of the data that are logged may be incomplete, such as visit duration as discussed below. As a result, conclusions based on this data may lead to unsound business decisions (Ferrini, Mohr, 2009).

Visit counts are also inaccurate because most Web analytics programs define a visit as a sequence of page requests from a unique visitor within a certain amount of time, usually 30 minutes. Counting visits in this manner is inaccurate because it relies on an arbitrary 30-minute timeframe to define a visit. Any visit longer than 30 minutes is counted as another visit. So, if a Website provides extensive information, or if a visitor is researching information on a Website for more than 30 minutes, visit counts will be inflated. (Ferrini, Mohr, 2009).

Although they did not encounter a specific problem, as hublo.com document, they focused on the accuracy of log files analysis, and gave a solution. As a remedy for such inaccuracy, they concluded that it is better to use page tagging combined with the usage of cookies.

Review Discussion

Although the paper discussed some of the issues affecting the data accuracy, for both log files and page tagging, it is noticed that it did not give a solution for a specific problem and how to counter it. Instead, in his ten recommendations he recommended "How to deal with data" or " What to log" points.

As we can notice, no research addressed all topics, which implies that no research gave a way to acquire hundred percent accuracy, which in theory and in practical (at least for the current days) is impossible.

Although in "Web Analytics" (Ferrini, Mohr,2009) the authors did a great work on explaining the accuracy issue, their weakness point was that they recommended page tagging which has its own weaknesses, since this method will be disabled if the user disabled JS in his browser.

In the other hand, in the document issued in the website (5 Threats to Data Accuracy, 2014) there was a solution for each problem it addressed; the solution for how not to count robots is very weak, because any unknown robot will not be counted. However, his solution to cross-device browsing is notable that can be developed further.

Overview of the study

Data web analytics, is a rapidly growing aspect of the web technology that is getting more important with time. Understanding different aspects of the data related to the website visitors, such as their location, time of visit, duration of stay, and knowing who is a new visitor, or a returning one, is of big importance and should be taken in consideration in order to improve the website.

However, it is not easy to collect accurate data, as there are many threats that can trick the collection of data, and give inaccurate readings, which may lead to incorrect decisions, some threats may prevent from collecting real unique users number, or knowing if it is human or robot that is requesting a page in the website.

We have used GA to test its reading while enumerating over some known and common threats, such as disabled JS, robot requesting page, dynamic IP addresses, and we have proposed some new ideas and tested it, and showed how it can help to overcome some threats, and increase the accuracy.

The main objective of the study was to enumerate the common threats to data web analytics, and propose solutions for them. To do this, the research methodology contained different steps: collecting common threats, analyzing threats, proposing new solutions, and in the last step explained the solutions in practical. In the first, the threats were collected from different resources, such as books and journals. We found that the two most common problems, are the inability to detect unique users at a high accuracy, and the confusion between the real users and robots, which were focused on when searching for solutions, it was kept in mind that the solutions should be simple yet reliable. Then a test was done to see how the current threats affect the readings of the metrics, different tests were done on different threats, in the last step the new solutions were tested and their result were shown in the results and discussion chapter, and some results were compared against the results of testing with the current threats.

Concluding remarks

The results were found to be helpful, as in our test we had the opportunity to track the user no matter from which IP he had used in requesting the website, also the system was able to distinguish between human requests and robot requests.

However, as it was noticed, some methods despite the reliability and the great help in increasing the accuracy of the analysis, they do have some limitations or weaknesses that should be studied and resolved, and so no method can give a full accuracy.

Problems and limitations

The main limitation faced in this study was in resources and information collection. The topic is somehow not very spread, and most of the studies conducted on it are very similar, and sometimes redundant, so we tried to collect as many references as it's possible in order to extract the useful information from each one.

Recommendations for future research

The thesis is one of the bus stops of this topic, which is a very vital topic in web technology. We can note some points that could be developed further and enhanced:

- The idea of using social media is depending on JS, which might be disabled by the user, so it is better to check for a way that cannot be interrupted by the user.
- Adding the RequestByUser variable, is beneficial, but one recommendation is to be sent in the HTTP header by default.
- Login to the browser is good, but it will be better if all the browsers have a common DB so one user can log in from any browser, and still be detected as one user, the question is: is it applicable or not?

References

1. Brian Clifton,2010, *Advanced Web metrics whitepaper: Understanding web analytics Accuracy*
2. *5 Threats to Data Web Analytics and How to Prevent them*,2014, Available from :
   < http://blog.hublo.com/2014/04/02/5-threats-to-data-accuracy-in-web-analytics-and-how-to-prevent-them/>
3. *Server Logs Hit Counter Web Analytics,* Available from:
   <http://www.theedifier.com/blogging-blogger/server-logs-hit-counter-web-statistics.php>
4. Anthony Ferrini,Jakki J. Mohr,*Uses, Limitations, and Trends in Web Analytics*,University of Montana, USA
5. *Cache,* Available from:
   http://en.wikipedia.org/wiki/Cache .
6. *Cookie Deletion White paper,* Available from:
   http://www.comscore.com/Press_Events/Presentations_WhitepaperAnalytics Accuracys/2007/Cookie_Deletion_Whitepaper
7. *Web Analytics,* Available from:
   http://www.webopedia.com/TERM/W/Web_analytics.html