Web crawlers on a health related portal: detection, characterisation and implications

Gawesh Jawaheer
City eHealth Research Centre
City University London

Patty Kostkova
City eHealth Research Centre
City University London

Abstract

Web crawlers are automated computer programs that visit websites in order to download their content. They are employed for non-malicious (search engine crawlers indexing websites) and malicious purposes (those breaching privacy by harvesting email addresses for unsolicited email promotion and spam databases). Whatever their usage, web crawlers need to be accurately identified in an analysis of the overall traffic to a website. Visits from web crawlers as well as from genuine users are recorded in the web server logs. In this paper, we analyse the web server logs of NRIC, a health related portal. We present the techniques used to identify malicious and non-malicious web crawlers from these logs, using a blacklist database and analysis of the characteristics of the online behaviour of malicious crawlers. We use visualisation to carry out sanity checks along the crawler removal process. We illustrate the use of these techniques using 3 months of web server logs from NRIC. We use a combination of visualisation and baseline measures from Google Analytics to demonstrate the efficacy of our techniques. Finally, we discuss the implications of our work on the analysis of the web traffic to a website using web server logs and on the interpretation of the results from such analysis.

Keywords: crawlers, visualisation, web analytics, web server logs.

1 Introduction

Web crawlers also known as crawlers, spiders, bots or robots are programs that visit websites in order to download their web pages. Crawlers are employed for non-malicious reasons such as indexing content for search engines [1] as well as for malicious reasons such as harvesting email addresses for spammers [2]. Typically, traffic from crawlers need to be identified and removed from the analysis of traffic to the web site.

In our work, we have analysed traffic to several health related web sites using their web server logs. Analysis of website traffic is essential to gain an understanding of its users; bespoke analysis of present and historic web server logs can answer the important questions of popularity and user behaviour. However, unidentified crawlers inflate the results of such analysis. And very often there is the tendency to overlook this issue as it is not in the interests of web hosting service providers to present less impressive data. The aims of our research work are to:

1. characterise malicious web crawlers,
2. develop heuristics for semi-automated identification of web crawlers based on a blacklist database, characteristics of their online behaviour and 3D visualisation of the web traffic and
3. analyse the performance of our crawler removal techniques.

In this paper, we describe our effort of detecting and removing crawlers visiting the National Resource for Infection Control (NRIC) portal, www.nric.org.uk. We discuss the techniques we used and illustrate those using examples from 3 months of web server logs we collected. We characterise the requests from crawlers by analysing their online behaviour [1, 2]. We use visualisation as part of the process to ascertain the efficacy of our techniques

In order to ensure a consistent understanding of the material in this paper, we provide definitions of the following terms and notations we used [5]:

- page view: a request made to the web server for a web page
- visitor: a user who views a web page on a web site and is identified by the IP address of his computer
- visit/session: sequence of requests from a uniquely identified client
- web site traffic/traffic: refers to all the visits to a web site

We start with an overview of the two main techniques of web site traffic analysis in Section 2. In Section 3, we present our work on crawler detection by describing the techniques we used. We proceed in Section 4 by illustrating those techniques using web server logs we collected from NRIC. In Section 5, we illustrate how visualisation can aid the crawler detection process and also ascertain the efficacy of our techniques. Finally, we present some related work and our conclusions in Sections 6 and 7.

2 Web site traffic analysis

Each time a resource hosted on a web server is requested on the web a record is written in the web server log. A resource can be an HTML file which is typically rendered on the user's web browser as a web page. Or it can be an image, a multimedia file or other media that can be accessed by a web browser or any other web client such as a crawler.

For each resource requested, the web server records the IP address of the web client, date and time of the request, the HTTP protocol used, the name of the resource, the HTTP protocol status of the server response, the size of the response, the identification of the web client and details of cookies used.

There are two means of analysing the traffic to a website; either by using web server logs or page tagging. The first method uses a software called a web server log analyser or just log analyser, to process the web server logs and report on various characteristics of the traffic such as the IP address of the web client, pages requested, frequency of requests, etc. The second method involves adding some lines of JavaScript code to 'tag' all pages that we want to analyse. Each time a tagged web page is accessed, the JavaScript code sends information about the access to a remote server which aggregates the data and reports on key parameters. Table 1 shows the various features of both techniques.

Table 1 Features of log analyser and page tagging techniques

	Log analyser	Page tagging
Requires changes to web pages	No	Yes
Analyse historical data	Yes	No
Records visits from crawlers	Yes	No
Requires web browser cooperation	No	Yes
Records requests for all content served by website	Yes	No
Records additional information without modifying URL	No	Yes
Records client side events	No	Yes
Web administrator owns the data	Yes	No
Records access to cached web pages	No	Yes

The main advantages of the using the log analyser are that the user owns the data, no changes are required to web pages, historical web server logs can be analysed and the analysis can be done with any log analyser software. The main disadvantages are that traffic to cached web pages cannot be recorded and traffic from crawlers are recorded and thus needs to be removed. However, as most of the web pages of the NRIC portal are dynamically generated, thus cannot be cached, the first limitation is not applicable to our work. In regards to the traffic from web crawlers, we will show that it is feasible to identify and eliminate the majority of web crawlers. Although, this identification and elimination process represents additional work, the benefits of the log analyser over page tagging more than compensate for it. Besides, being able to record traffic from web crawlers provides valuable information for carrying out Search Engine Optimisation (SEO) on the website. By analysing the content of the web pages requested by the crawlers of a search engine and the frequency of their visits, one can optimise the visibility of the website's pages within the search results of that search engine. Unfortunately, this discussion is outside the scope of this paper. In the next section, we describe the work we have done on the detection and elimination of web crawler traffic.

3 Detection

All non-malicious crawlers adhere to the 'Robot Exclusion Protocol' [1]. This is a convention which allows the administrator of a website to prevent non-malicious crawlers from accessing part of an otherwise public website. It is implemented in the form of instructions provided to crawlers in a text file, conventionally named robots.txt. Hence, all non-malicious crawlers will try to find the robots.txt file on the web server, download and interpret it. This makes the detection of non-malicious crawlers as trivial as identifying the IP addresses of records in the web server logs where one of the requested resources is the 'robots.txt' file. Also, non-malicious crawlers will identify themselves to the web server by setting their identification details in the User-Agent field of the HTTP protocol and ensure those details are different compared to those from web browsers.

On the other hand, malicious crawlers will use all subterfuges to hide their trails [4]. For example, they will falsely identify themselves as genuine web browsers and they will not request the robots.txt file. Thus, detection of malicious crawlers is a challenge. There is no single technique that will identify all crawlers in a web server log. Instead, there is a combination of heuristics and techniques that can identify the majority of crawlers. Thus, we describe several of these techniques and heuristics next.

3.1 Characterisation of crawlers by their online behaviour

A crawler accessing a web site exhibits a different online behaviour to a user accessing a web site using a web browser [4]. Crawlers download all the resources on a web server indiscriminately whereas web browsers

download only a subset of resources, most likely to be resources related to each other. Requests from web browsers are limited by the physical ability of the human body whereas requests from crawlers are programmatically defined, automated and often use very sophisticated algorithms to avoid detection. Malicious crawlers try to mask themselves as web clients by mimicking the human behaviour. In order to hide the indiscriminate nature of their requests, malicious crawlers distribute their requests over time and use different IP addresses. Furthermore, malicious crawlers can mimic the limitations of the human body by reducing the request rate and adding some random pauses between requests.

Analysis of the online behaviour of malicious crawlers is a very time consuming process. As such it is used to identify crawlers in the beginning of the page request distribution where accesses from crawlers easily stand out from genuine traffic. We worked out a set of heuristics that help us semi-manually identify crawlers through analysis of their online behaviour:

- Absence of referrers, especially when subsequent pages requested are hyperlinked in a previously downloaded page
- Indiscriminate downloading of content
- Sequential downloading of content without a referrer page
- High web page request rate, especially in the case of pages not referred from another page.
- Lack of diurnal pattern of activity
- Regularity of requests: same time and same number

Identifying these characteristics of crawler behaviour, we developed the techniques described in the next two sections.

3.2 Database of crawlers

Over time we built a blacklist database of crawlers with the IP address as key identifier. The crawlers detected through online behaviour analysis are added to the database after each analysis run. This database is then used to identify repeat offenders. Using the database makes the inherent assumption that once an IP address is used by a crawler it is not used by anything else afterwards. This is not true as malicious crawlers can sometimes be running on personal computers that have been hacked or hijacked. In this case, the IP address of that personal computer will be forever blacklisted. So, firstly we identify the non-malicious crawlers using online behaviour analysis as described previously in Section 3.1. We then update the database using the identified crawlers. We process the logs a second time and remove the crawlers those IP addresses appear in the database. Thus, this leaves us with malicious crawlers which are accessing our web server for the first time or which has such low level of activity that they are in the long tail of the distribution of accesses and thus very hard to identify. We employ the online behaviour analysis described in Section 3.1 to identify such users. However, as this is a very time consuming step, we apply it only on crawlers at the beginning of the distribution as they have a bigger impact on the accuracy of the analysis. Using the heuristics in Section 3.1, we analyse the online behaviour of the first 20 web clients that have the highest number of requests to the server. This is done on a monthly basis and the cut off point may depend on the monthly level of traffic. In our case, we found that 20 provides a good balance between timeliness and effectiveness.

By identifying non-malicious crawlers in the web server logs using the database in the first place, the time consuming technique of online behaviour analysis is reserved mainly for malicious crawlers, thus saving time.

4 Experiment Results

Using the techniques discussed above allowed us to accurately evaluate the traffic to our medical portals (NeLI, NRIC) over a number of years using log analysers. However, to justify the manual effort we ran an experiment using web server logs from the NRIC portal collected over 3 months, from 1st March 2010 to 31st May 2010 against page tagging data collected by Google Analytics for the same period. We used a period of three months as we only had three months of data collected by Google Analytics as shown in Table 2. In contrast, we have 5 years of web server logs for NRIC.

Table 2 Crawler-free figures for NRIC from Google Analytics

Month	Page Views	Visitors
Mar	16,966	6,015
Apr	14,479	4,793
May	16,480	4,916

The objective of this experiment is to demonstrate the ability and assess the performance of our techniques in identifying web crawlers from the logs. We compared our results obtained using the Sawmill Log Analyser against the reports generated from Google Analytics. As explained in Section 2, page tagging services such as Google Analytics do not record crawler traffic; thus provided the base line for our experiments.

However, there is a limitation to such a comparison. Google Analytics does not show the IP addresses of the web browsers. Thus, although the figures from Sawmill and Google Analytics may agree on the number of page views or visitors, we cannot tell if the traffic measured in Sawmill is actually from a crawler and not a web browser. But we will show in Section 5, how we have used visualisation to ascertain this. In the next sections, we

illustrate our crawler detection techniques using the results from the experiment.

4.1 Step One – removal of crawlers using blacklist database

Table 3 shows the monthly page views and visitors produced by the log analyser, Sawmill, after crawlers have been removed using our blacklist database. Also, non-malicious crawlers were identified through their requests for Robots.txt file and their identification details in the User-Agent field and the blacklist database updated accordingly such that at the end of this first step of the detection, the database contained the IP addresses of 13,613 crawlers.

Table 3 Figures from log analyser after crawlers' removal using blacklist database

Month	Page Views	Visitors	% Page views difference	% Visitors difference
Mar	20,795	6,036	+22.6	+0.3
Apr	15,192	4,645	+4.9	-3.1
May	16,986	4,793	+3.1	-2.5

The percentage differences columns in Table 3, show the relative difference of the relevant figures from our log analyser against Google Analytics. As shown by the low values in the %Visitors difference', solely using the database of crawlers can identify the majority of crawlers. The negative difference in April shows that we may have been too aggressive in building our database and as a result we have some false positives in our blacklisted Visitors. Nevertheless, we observe in Mar, that the small difference in visitors may still lead to big discrepancies in the number of page views. This is because crawlers typically access higher number of web pages than genuine web browsers and even missing a few ones in the identification will cause large discrepancies in the number of page views.

4.2 Step Two – removal of crawlers through analysis of the online behaviour

As mentioned in Section 3.1, in practice it is not feasible to do an analysis of the online behaviour for all the visitors. Hence, in this step, we sort the visitors in terms of descending order of page views and carry out the analysis of the online behaviour for the Top 20 visitors. Through our experiments, we found that 20 gave a good balance between efficiency and efficacy. The finding in Section 4.1, which showed that a small no of visitors cause a large discrepancy, validates our approach of maximising accuracy of detection against speed of detection by focussing on the visitors that cause the highest traffic. Figure 1, shows the Top 20 visitors to NRIC in March.

A quick glance at the number of page views shows that hosts 1 and 2 stand out from the rest and are thus good candidates for 'potential crawlers'. Nevertheless, we indiscriminately analyse the online behaviour of the Top 20 visitors shown in Figure 1, and identified 2 likely crawlers. In fact, host 1 turned out to be a genuine user (in fact likely to be many users using a corporate network infrastructure which appears as only one IP address). We identified two crawlers from the list of 20, namely hosts in positions 2 and 4. We repeated this step for visitors in April and May, identifying 5 crawlers in each; positions 4, 10, 18, 19 and 20 for April's logs and positions 1, 6, 10, 14 and 16 from May's logs. Overall, we analysed requests from 50 unique IP addresses in three months and identified 10 crawlers (1 crawler was repeatedly identified in all three months), i.e 20% of unique IP addresses analysed were crawlers. In these three months, NRIC received requests from 14,433 unique IP addresses. Thus, the crawlers identified in this step represent only 0.07% of the visitors. We added the 10 crawlers to our blacklist database. Thus, Table 4, shows the new figures after Step Two.

Table 4 Figures from log analyser after crawler removal using analysis of online behaviour

Month	Page Views	Visitors	% Difference Page views	% Difference Visitors
Mar	19,329	6,032	+13.9	+0.3
Apr	14,828	4,640	+2.4	-3.2
May	16,038	4,787	-2.7	-2.6

Step Two improved the accuracy of the figures from the log analyser compared to Google Analytics for the first two months. May's figures went down, most likely due to some false positives. As we discussed in Section 4, not having the IP addresses in Google Analytics mean that we need another way of confirming that we have reduced the number of crawlers rather than removing genuine web browsers. The visualisation technique we describe next helps us do that.

5 Visualisation

Despite all our best efforts, some crawlers may still remain undetected. This is particularly true of new crawlers which may not exist on the blacklist database or crawlers with such low level activity that they remain hidden in the long tail of the distribution of web server requests or of crawlers which employ the ultimate subterfuge of mimicking humans' access characteristics. Furthermore, we need a way to verify that we are actually

removing crawlers and not leaving crawlers but removing web browsers. A very simple and quick to implement technique which we found very effective is 3D visualisation of the requests.

	Hostname	Hits	0 - 100 %	Page views
1	194.176.105.46	0	0.0 %	1,441
2	91.205.96.13	0	0.0 %	1,272
3	193.60.159.61	0	0.0 %	212
4	146.101.163.36	0	0.0 %	134
5	86.20.38.46	0	0.0 %	129
6	86.159.211.87	0	0.0 %	121
7	92.2.125.204	0	0.0 %	112
8	92.11.136.84	0	0.0 %	100
9	59.90.27.246	0	0.0 %	92
10	195.89.26.192	0	0.0 %	85
11	81.145.165.2	0	0.0 %	81
12	195.89.27.224	0	0.0 %	79
13	81.147.70.71	0	0.0 %	75
14	79.72.154.194	0	0.0 %	73
15	86.184.226.143	0	0.0 %	72
16	77.68.120.191	0	0.0 %	67
17	86.133.193.225	0	0.0 %	63
18	81.153.153.182	0	0.0 %	62
19	62.6.175.161	0	0.0 %	56
20	122.166.130.120	0	0.0 %	55

Figure 1 Top 20 visitors to NRIC in March 2010

Figure 2 shows a 3D plot of the NRIC web traffic including traffic from the crawlers. The x-axis represents the IP addresses which are converted to integers and ordered. The z-axis represents 'Content', i.e., the requested resources. Content is represented by an integer and ordered according to popularity, with the most popular requested resource (usually the home page) having an integer value of 1. The y-axis represents date/time of requests.

Figure 2, shows that the majority of requests are for the most popular resources on the web site. This is representative of web traffic and most web sites exhibit such characteristic. On the other hand, note the continuous vertical trails that cut across most of the content space. These represent traffic from crawlers and there are many of them. In contrast, in Figure 3, which shows a plot of web traffic after Step One of the crawler removal process, there are only a few continuous vertical trails (3 of them are visible from the angle shown). This represents visual confirmation of the efficacy of Step One of the crawler removal process. Step Two produces a further reduction in the number of crawlers as shown in Figure 4 where only one vertical trail is barely visible.

Visualisation is a recognised technique for displaying large amounts of data in a concise manner in such a way that one can easily and quickly assimilate the relationships between the parameters of the data. It is an important tool that can be used throughout this semi-automated crawler removal process. The 3D plots of web traffic provide a quick and easy assessment of the efficacy of the presented techniques. The plots can also help improve the accuracy of the techniques by visualising traffic where the analysis of online behaviour was unsuccessful in clearly identifying between crawler or browser.

.

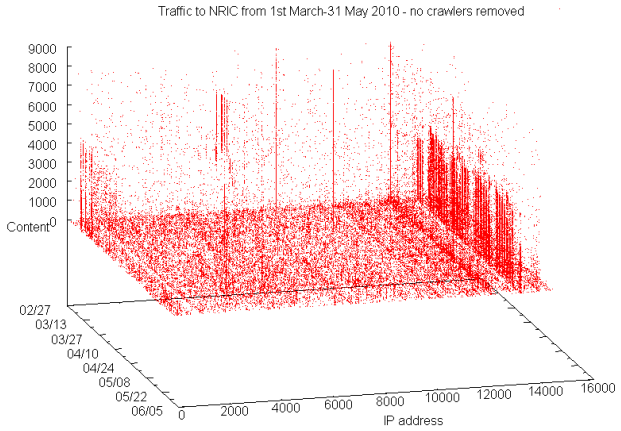

Figure 2 Visualisation of NRIC traffic without any crawler removal

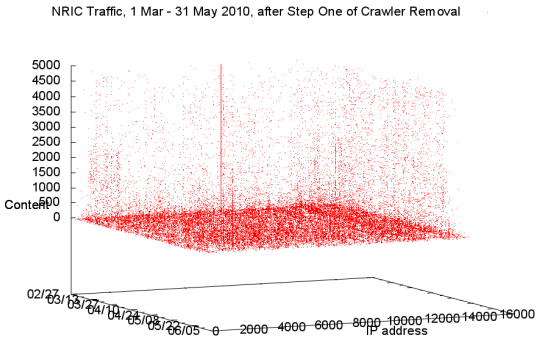

Figure 3 Visualisation of NRIC traffic after Step One in Crawler Removal

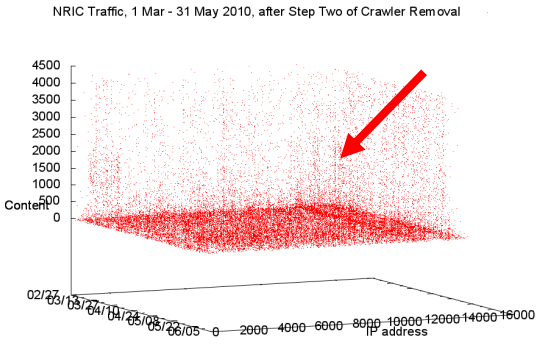

Figure 4 Visualisation of NRIC traffic after Step Two in Crawler Removal

The other benefit is the 3D nature of the plots and the ability to rotate them along the three axes. The resulting changes in the point of view of the plotted data can reveal crawlers which were previously hidden. For example, crawlers which download content in chunks over several days will appear as several broken vertical trails at all other angles except when viewed at an angle where the date/time axis is perpendicular to the screen

6 Related Work

Crawler detection has been studied before. In their work [4], the authors discuss the issues involved in identifying crawlers in an online journal. First, they used techniques that fall within Step One of our crawler removal process; identification through visits to the robots.txt file, details on the DNS entry, USER-AGENT field in the HTTP request and an external database of known crawlers. They estimate that at the end of their equivalent Step One, they identified 32.6% of requests were from crawlers. Their equivalent Step Two involves analysing the online behaviour of crawlers to identify metrics that can aid crawler detection. A further 20% of traffic is identified as being from crawlers after Step Two. They discuss the increasing prevalence of crawlers that hide their activity by mimicking human like browsing and the difficulty this causes in identification. They do not perform the visualisation step which helps us check the efficacy of our techniques. Their crawler detection is geared towards a fully automated approach whilst ours is semi-automated. We feel that the latter approach gives a better balance between efficacy, efficiency and accuracy.

The visualisation of web traffic as 3D plots such that they can help identify crawlers was proposed by [6] in an online article. We customised the Perl scripts provided by the author to generate the 3D plots.

7 Conclusions

Evaluating the traffic to a web site is an essential task in better understanding its usage and popularity. This is of intrinsic value for web site marketing and better analysing the online user needs, navigation behaviour and overall traffic. Using log analysers present significant benefits compared to page tagging tools. However, they bring the complexity of identifying and eliminating traffic from crawlers in the web server logs. Through our investigation of the characteristics of the online behaviour of web crawlers, we have developed a series of techniques to eliminate crawlers. We have shown that these techniques do significantly reduce the number of crawlers to the extent that the results from the log analysers are on par with those of page tagging tools.

Many website providers are interested in results of web traffic reports generated from log analysers but do not understand the procedure required for identifying web crawlers and often are unaware the results are typically over-optimistic. And not accounting for even a few web crawlers can cause large discrepancies in the figures for page views.

Our experiment has also shown that the techniques discussed in this paper can produce false positives. In our future work, we are investigating techniques for identifying these false positives.

8 References

[1] Y. Sun, Z. Zhuang, and C.L. Giles, "A large-scale study of robots.txt," Proceedings of the 16th international conference on World Wide Web - WWW '07, 2007, p. 1123.

[2] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage, "Spamalytics: An empirical analysis of spam marketing conversion," Proceedings of the 15th ACM conference on Computer and communications security, ACM, 2008, p. 3–14.

[3] M.D. Dikaiakos, A. Stassopoulou, and L. Papageorgiou, "An investigation of web crawler behavior: characterization and metrics," Computer Communications, vol. 28, 2005, pp. 880-897.

[4] P. Huntington, D. Nicholas, and H.R. Jamali, "Web robot detection in the scholarly information environment," Journal of Information Science, vol. 34, 2008, pp. 726-741.

[5] "Web analytics," Wikipedia, 2010.

[6] R. Varghese, "A New Visualization for Web Server Logs," O'Reilly Sys Admin, 2010.