Parsing and filtering of web log data and data analysis

October 22, 2015

1 Prescribed queries

1.1 The top ten files/pages requested

The following table shows the most common requests received by the server. The top request appears to be for a CSS library (or similar) so should probably be disregarded. The homepage appears to be 3rd in the list of the remaining requests, possibly suggesting that traffic is being linked to externally more than people are hitting the homepage.

n	path
1324	/library/conditionalstyle.asp
957	/uk/letters/letters.asp
872	/uk/home/Default.asp
631	/Default.asp
511	/uk/financialcentre/tax_calculator_tool.asp
471	/uk/letters/default.asp
388	/uk/letters/resignation_letter_generator_form_v2.asp
387	/uk/discussion/new_topic.asp
281	/uk/letters/resignation_letter_generator_generate_v2.asp
206	/uk/financialcentre/tax_calculator.asp

1.2 The top ten IP addresses (or users) who requested the most URLs.

From the data provided, there was never any username set so the username was disregarded in the query. The following shows the top 10 IPs.

n	IP
554	195.149.39.85
328	65.214.36.156
194	65.214.36.152
192	213.199.149.236
88	209.140.222.149
84	195.92.168.177
79	62.254.0.7
72	192.168.1.6
60	12.47.98.180
58	62.255.64.5

1.3 The top three most active hours (most requests per hour).

The following table shows the three busiest hours.

n	hour
645	20
608	14
597	18

1.4 The number of requests per query method.

The majority of requests were GET requests which would be expected from a WWW webserver. The 10 that had no method type were also errors of various types.

n	method
9583	GET
240	POST
10	HEAD
10	-

2 Additional queries

2.1 Errors served with internal referrals

The following table shows error documents that were served from a visitor following an internal link. This would be of interest to the webmasters/content owners/developers to identify potential faults with the content or application server.

n	path	status	referrer
6	/uk/letters/letters/letterform.asp	404	http://www.i-resign.com/uk/letters/kissmyass_resign.asp
6	/uk/letters/workinglife/viewarticle_4.asp	404	http://www.i-resign.com/uk/letters/dilbert_resign.asp
3	/uk/letters/halloffame/	404	http://www.i-resign.com/uk/letters/kissmyass_resign.asp
3	/us/financialcenter/_tc.asp	500	http://www.i-resign.com/us/financialcenter/federal_tax_estimator_2.asp#form
3	/uk/letters/letters/http:/www.i-resign.com	404	http://www.i-resign.com/uk/letters/letters/letterform.asp
2	/uk/letters/resignation_letter_generator_generate_v2.asp	500	http://www.i-resign.com/uk/letters/resignation_letter_generator_generate_v2.asp?country=&RT=n
2	/uk/stress/	404	http://www.i-resign.com/uk/legaladvice/top.asp
2	/uk/letters/letters/http:/http:/www.i-resign.com	404	http://www.i-resign.com/uk/letters/letters/http%3A//www.i-resign.com
1	/cgi-bin/formmail.pl	404	http://www.i-resign.com/
1	/cgi-bin/formmail.cgi	404	http://www.i-resign.com/
1	/uk/discussion/new_topic.asp	500	http://www.i-resign.co.uk/uk/search/Default.asp?free=&query=holiday+pay&p=1

2.2 Traffic from Search Engines

In Januray 2003, according to OneStat, the top three search engines were Google, Yahoo and MSN with 54.7%, 22.1% and 9.5% respectively. The following table shows the number of requests from each of those engines. They are not shown as a % as that would be misleading (not all search engines were considered so the traffic would be skewed) but it can be seen that they are similar in ratio to the reported statistics of the time.

n	**search_engine**
407	Google
181	MSN
300	Yahoo

3 Appendix

3.1 Queries that would be interesting but no longer relevent

It would have been interesting to compare the IP addresses with Geo location datasets to analyse where traffic is coming from. It would also have been interesting to query PTR records for the IPs to more accurately identify crawlers and bots and also get a snapshot of ISPs. Unfortunately the current IP data would be far out of date now.