informationpower

IRUS download data – identifying unusual usage

Report by Information Power Ltd

July 2013

Contents

1. Task

“To test the feasibility of devising a set of algorithms that would ‘dynamically’ identify and filter out unusual usage/robot activity.”

2. Data supplied

a. Daily log files for the period July 2012 – mid-April 2013, with lines of data values based on the Tracker Protocol. Each line has key/value pairs identifying the client IP, client User Agent, date-time of the download event, repository used and identification code of document downloaded.

b. Prototype filters for User Agent and blacklist IP Addresses.

3. Previous Relevant Research

Analysis of log files is fairly standard practice, the main difference is in the type of service being analysed and the users being served.

> Web robot detection based on pattern-matching technique *Journal of Information Science April 2012 38: 118-126* Shinil Kwon, Young-Gab Kim and Sungdeok Cha
>
> Search log analysis: What it is, what’s been done, how to do it *Library & Information Science Research 28 (2006) 407-432* Bernard J. Jansen

Both studies were concerned with identifying types of data which can be used to differentiate between different kinds of user behaviour. They also come up quite a few problems. Whilst it easy to identify the major search engines, link checkers and other welcome visitors to the web site, others (spammers etc) try, by a variety of means, to behave just like regular users. The more refined the filtering system the more likely that regular users will be excluded along with the unwelcome ones, the same trade-off between recall and precision which confronts any search for information.

Unlike the files used in these and similar studies, the IRUS log files are not true log files as the interactions which lead to specific downloads are missing, so some of the potential evidence for distinguishing behaviour is absent.

There are various web sites with databases and search facilities to try to check if User Agent strings or IP addresses are known to belong to search engines, spiders, and bots of various kinds. None of these are necessarily up to date or comprehensive and none offer an interactive service for automatic checking. These are best used once a set of suspect IP addresses or User Agents have been identified.

4. Initial Study of the IRUS Log File Data

There were 290 files; each representing a day's downloads from the repositories participating in the IRUS user study, with each file having some 100,000 records. To see what could be learnt from a day's records, a small database was created from them so that filtering of known search engines and bots could begin and others identified so that analysis of genuine user behaviour could begin. The first pass of the filtering removed 68% of the records, and after adding additional obvious terms to the User Agent filter 85% of the records were removed. After this the records were organized on the basis of 'client' activity. Records were combined on the basis of req_id/req_dat/rfr_ip sorted by url_tim to give all the downloads for the day by a particular 'client', along with a number of metrics for example: total duration of activity, the number of downloads in a minute, time between each download and total number of downloads in day. These metrics were used to create a number of rules with thresholds to flag up 'clients' whose behaviour looked suspicious and needed checking out.

A lot of activity in a short period of time seems suspicious, one client - who turned out to be a spammer of some kind - had 500 downloads of the same document in a minute. Randomly checking a dozen or so of the IPs, with a short duration of activity and a limited number of downloads (10), produced some known spammers; one of which used multiple IP addresses with only one download each per day. At low levels of downloads per day the metrics and rules used cannot reliably separate out acceptable usage from the spammers and robots.

5. White List

An alternative approach was adopted to see if IPs belonging to know academic institutions could be identified in the download data logs. A list of IPs addresses belonging to a number of UK academic institutions was matched against the residue of filtered records; around 18% of the IPs were matched accounting for 16% of the downloads. (The Cranfield hashed IP addresses could not be used for this.) The arXiv repository reports usage from major academic institutions along with other more general usage data. They either have a list of the IP ranges of these institutions or are using reverse IP look up services to identify the owning institutions. Knowing the institution behind the IP does not guarantee that the usage is acceptable!

6. Using all the Data

Having an identity for some of the academic clients showed a pattern of behaviour which looked like it might be more useful for separating the sheep from the wolves. Frequently the identified IPs had multiple User Agents associated with them; in addition the amount of duplicate downloading of the same document was much lower than identified spammers. It was also apparent that data from one day was not going to adequately quantify behaviours. The next set of rules looked at IP usage backwards one month and then across the entire period. Suspicious usage was confirmed in some cases but in others it turned out the IP was only active for a few days. In one case an Iranian university IP downloaded 200 identical documents in one day and did show up again. This appeared to be a pattern which suggests that the Open Access Repositories having no form of access control attracts the attention of spammers etc but they see no value in returning. Only time will tell if this is true.

7. Duplicate Downloads

One puzzling aspect of the download data is the volume of the same document being down loaded multiple times by the same client at more or less the same time. This is not due to accidental double clicking or impatient users frustrated by poor performance. One possibility is the use by clients of ‘download managers’ which will download in parallel segments of a file to get faster download times. This does not explain the apparent downloading of hundreds of copies. Cutting out the duplicate downloads will have a significant impact on the final filtered usage statistics.

8. Rule Parameters

The rules are part of a process which starts with filtering out records which match previous identified criteria for unacceptable behaviour in relation to usage statistics. The rules are there to flag up behavior which looks suspicious and needs researching and perhaps adding to the filter criteria.

There is no hard and fast description of either suspicious or acceptable behaviour; it can’t necessarily be judged on the basis of one day’s usage records or a month’s. [A month was selected because COUNTER users statistics are reported monthly.]

An ip address can be used by many individual users over the day, but with the data in the logs there is no way of knowing when one starts and finishes, in fact several individuals could be using the ip address at the same time, with different or the same user agent using the same or different repositories, what is missing from the logs is the port number which would give a closer match to a ‘user’, though over the course of a day even this will probably reflects usage by more than one individual.

The focus for the rules is the use per ip address, if there is more than one user agent associated with it this suggests more ‘individual’ users and their usage can be spread over several repositories at the same time.

8.1 Metrics

Day metrics

$day_hits=($hit_count/$rfr_ip_count)/$agent_count; # total hits per repository per user agent during day

$day_hit_level=($dist_hit_count/$rfr_ip_count)/$agent_count; # distinct hit per repository per user agent

Month metrics

$month_hits=($sum_hits/$num)/$agent_count_range; total hits per repository per user agent during month

$month_hit_level=($max_dist/$num)/$agent_count_range# distinct hit per repository per user agent

Rules using the metrics, each rule has an id so that output results can be linked to the rule and the thresholds altered if necessary

```
if ( ($month_hits > 100) and $month_hit_level > 40) {$flag='Check 1'} # lot of use on day and in earlier month

elsif ( ($day_hits > 10) and $month_hit_level > 10) {$flag='Check 2'} # a day’s use may  be more significant than a month’s

elsif($month_hits > 100 and ($sum_dist_hits/$sum_hits < .2)) {$flag='Check 3'} # high proportion of hits with the same id}

            else {

if (($day_hits > 10 ) and $day_hit_level > 20) {$flag='Check 4'} # medium number of unique hits with same id

elsif($day_hits > 10) and ($day_hit_level/$day_hits < .2)) {$flag='Check 5'} # low number of hits with same id}}
```

8.2 Final Steps

Flagged ip addresses are checked against the white list to see if it is usage by a ‘known’ institution and can be accepted. It might have been flagged because of unusually high usage caused by tests. If the flagged address is not known, the use throughout the whole database of accumulated log records in terms of total and distinct downloads is added to the output.

Checking the flagged ip addresses

All the filters are based on empirical evidence as are the thresholds. IP and usage agents can be searched for with search engines looking to see if they have been reported as spammers or other kinds of unwanted visitors. If they have been then the IP address or user agent can be added to the appropriate filter. Unfortunately this is hit or miss and there can be no formal identification of many suspicious ip addresses. Because of this there may be little point in checking suspicious usage if it only occurs for a day or so or over the year usage is pretty low.

9. Conclusion

Unusual usage identified by the rules needs following up to see if the IP address or the Usage Agent need adding to the filters. Often the source of the usage will not identifiable or recognized by searching on the web. If the overall historic usage is low or the short lived the best thing to do would be to keep a record of the suspicious IP and see if it turns up again. This means that there will be a proportion of usage that looks suspicious but might not be. On the other hand there will certainly be an unidentified proportion which is suspicious but not recognizably so.