IRUS-UK position statement on the treatment of robots and unusual usage

Version 1 November 2013

The processing of repository downloads, which IRUS-UK has done to date, is based upon and conforms to the COUNTER-PIRUS Code of Practice (http://www.projectcounter.org/documents/Pirus_cop_OCT2013.pdf). COUNTER provides a list of well-known robots, whose usage should be removed as a bare minimum. The list is used as part of the audit process and is not intended to be a comprehensive list. The need for more sophisticated rules and processes is well understood.

IRUS-UK has, from the outset, added further filters to remove more user agents identified as robots and applied a simple threshold for 'overactive' IP addresses which eliminates:

- All downloads from IP addresses where there are more than 200 downloads in a day from a repository - except for known proxy servers
- Almost all downloads from IP addresses where there are more than 100 downloads in a day from a repository – except for known proxy servers and depending on the pattern of usage

These filters remove most of the 'big hitters', but there is more that can be done to refine and improve the statistics.

As an important step towards removing usage based on observed behaviours, we commissioned a report into the establishment of an adaptive filtering system. It is based on weightings and thresholds, in order to more effectively identify unusual and unacceptable usage patterns. A copy of this study, conducted by Information Power Ltd, is available on our website (http://www.irus.mimas.ac.uk/news/IRUS_download_data_Final_report.pdf).

We are actively working on applying thresholds to raw downloaded data and now have 12 months' worth of data to analyse. We continue researching and experimenting in this area but, we need to find a balance whereby we are able to exclude previously unidentified robots and abnormal usage within realistic timescales and budgets. We feel that a community-driven approach would be the most pragmatic and beneficial.

We are working in collaboration with COUNTER and the results of this work will inform future releases of the COUNTER-PIRUS Code of Practice. (Interestingly, one of COUNTER's auditors expressed a concern that adaptive filtering could potentially prove so complex that it would increase the costs of COUNTER audits.)

Following our work on applying adaptive filtering we will refine the data ingest process, restate the data and make both old and new usage data available so that you can compare the two and see the improvements we have made.

We also need to bear in mind that automated downloads are not necessarily robots, e.g. an institution doing a major literature search uses a script to cross search a number of databases and repositories – should this be excluded?

Underlying all this is the question as to what actually constitutes genuine usage, e.g. a lecturer with a class of 30 students in a computer lab asks them all to practice downloading the same three items – is this genuine usage?

In IRUS-UK, and most other statistical packages, we are using downloads as a proxy for usage, i.e. we are making a qualitative statement using a quantitative measure – this can never be totally accurate

Thus, all measurements are ultimately a judgement call and necessarily arbitrary. What distinguishes IRUS-UK from other statistics packages is that we work to a transparent, global standard that is applied consistently across all participating repositories.

Each available statistics package is set up to measure particular things in particular ways and we have no way of knowing how they have been set up. Therefore you are never going to get comparable figures. We have investigated a number of apparent discrepancies and found a number of different reasons for the variations in figures. In one case we found that a particular package excluded departmental proxies although these are not robots. We recommend that, in order to get a complete picture of all aspects of your repository's usage, you consider each package's strengths and use other packages, in addition to IRUS-UK, to provide those additional metrics that are not available in IRUS-UK, e.g. geographical coverage.