Developing COUNTER standards to measure the use of Open Access resources

Joseph W. Greene
University College Dublin and COUNTER Robots Working Group

Abstract: There are currently no standards for measuring the use of open digital content, including cultural heritage materials, research data, institutional repositories and open access journals. Such standards would enable libraries and publishers that invest in open digital infrastructure to make evidence-based decisions and demonstrate the return on this investment. The most closely related standard, the COUNTER Code of Practice (CoP), was designed for subscription access e-resources and ensures that publishers provide consistent, credible and comparable usage data. In the open environment, computer programs known as web robots constantly download open content and must be filtered out of usage statistics. The COUNTER Robots Working Group has recently been formed to address this problem and to recommend robot detection techniques that are accurate, applicable and feasible for any provider of open content. Once accepted, they will be incorporated into the COUNTER CoP 5. In this paper we describe the overall goals of the analysis, the scope and techniques for building the dataset and the robot detection techniques under investigation.

Keywords: Open Access, institutional repositories, OA journals, usage statistics, downloads, web robots, data mining

1. Introduction

Project COUNTER has brought together publishers, vendors and librarians to develop and maintain the COUNTER Code of Practice (CoP) since 2003. The adoption of the CoP ensures that publishers and vendors can provide consistent, credible and comparable usage data for e-resources including peer reviewed journal articles and e-books (COUNTER, 2017a). However, the COUNTER CoP was designed for resources that are only accessed behind subscription barriers. In the open environment, publishers and libraries offer free, unrestricted access not only to their designated communities, but also to users of computer programs, known as web robots, designed to automatically crawl the web for content. Web robots account for a very large percentage (between 40%

Received: 31.3.2017 Accepted: 21.5.2017 ISSN 2241-1925

and 85%) of open content usage and must be filtered out of the statistics for them to be meaningful (Greene, 2016, Huntington et al., 2008).

This paper describes the methodology used to test the effectiveness of a set of filters to detect and remove robot activity from open access usage statistics, balancing the highest accuracy of usage statistics with lowest barrier to implementation. The results of this research will inform and, pending community approval, become a part of the COUNTER Code of Practice 5. It should be noted that the research itself is currently in progress and is subject to variation from what is documented here.

2. Raw data to structured data

In September 2016, Project COUNTER formed a Robots Working Group, composed of volunteer experts from large publishers and vendors (EBSCO, Elsevier, Wiley) and representatives of smaller publishers, open access journal hosts, institutional repositories (IRUS-UK, DSpace, Eprints, Digital Commons) and aggregators (OpenAIRE). Three members offered the use of a subset of their data: Bielefeld University, which hosts a number of open access journals, IRUS-UK, which aggregates usage data from 127 open access institutional repositories (IRUS-UK, 2017) and Wiley, a large academic publisher that hosts a wide range of content, both open and subscription based. Each dataset included raw usage event data for the one week period of 3 to 9 October 2016.

The Bielefeld dataset consisted of Apache log files (anonymised via the final 2 octets of the IP addresses) from 7 open access journals, each hosted on PKP's Open Journal Systems (OJS). Three of the largest log files were selected (representing three journals). The IRUS-UK dataset contained item downloads from 97 UK open access institutional repositories. The Wiley dataset included all usage events on the platform, with registered crawlers including Googlebot and some technical partners removed. Characteristics of the datasets are described in Table 1.

Source	Original format	Lines, raw data	Lines, downloads only (*N*)	Sample size (*n*)	Confidence
Bielefeld OJS journals	Apache server logs	232,944	14,536	202	95%
IRUS-UK	TSV	1,935,689	1,935,683	204	95%
Wiley	TSV	30,419,834	5,098,763	204	95%[1]

[1]Some registered crawlers were removed from Wiley's raw data and the robots-to-total value is unknown, so confidence level may vary in the final analysis

Table 1. Data sources, sizes and samples

The datasets were converted into a usable format and imported into a PostgreSQL database. Each dataset presented its own challenges in terms of the restructuring. The OJS dataset included all HTTP requests to each site, so requests for full text downloads had to be accurately identified and extracted using regular expressions. The Apache combined format log file was then converted to SQL statements using regular expressions in order to import them into PostgreSQL. The Wiley dataset was quite large at 6.3GB and also included many events that were not downloads, which were removed. The final database contained 7,048,991 rows, each row containing data about a unique download event including data source, IP address, user agent, referring URL, request URL, unique item identifier, date, time and session ID (for Wiley items).

3. Robot detection and filtration experiments

With the data from each of the three sources in a single database, the downloads could be queried and sampled in a systematic manner. We devised a simple random sample calculator based on Sheaffer et al. (2006), assuming a robots-to-total ratio (*p* value) of .85, based on Greene (2016) and Information Power Ltd. (2013) for a confidence level of 95%. It should be noted that the confidence level for the Wiley sample may vary from this in the final analysis for two reasons: the non-inclusion of registered crawlers mentioned above, and the fact that the robots-to-total ratio likely differs from a fully open access dataset. In all, the sample totalled 610 downloads. The sizes of the individual samples are given in Table 1.

Additional variables, described in Table 2, were added to the sampled data. Wiley items were labelled open or closed access; 45% of the items in the download sample were found to be open access. A series of three passes were then taken through the dataset in order to determine whether they were made by humans or robots, following closely the methodology described in Greene (2016).

Field	Use
download_id	Unique identifier for the download event
session_id	Web site session ID (Wiley only)
item_id	Unique identifier of the item downloaded
IP	IP address that downloaded the item
origin	Organisation attributed to IP address
indicator	Indicator or sum of indicators used to determine robots[1]
other_indicators	Description of any other indicators about robot/human behaviour
counter_list	Boolean true where agent name would match against the COUNTER list of robots, crawlers and spiders

robot	Boolean representing if an item was manually determined to be a robot (1 = true = robot)
date	Date of the download
time	Time of the download
source	Source of the download data (IRUS-UK, Bielefeld, Wiley)
agent	User agent string provided for the download event
agent notes	Count of agents used by the IP over the course of the sample period; other info as needed
dl_peak_this_item	Total downloads of this item by this IP address during the period
dl_peak_any_item	Highest total downloads of any single item by this IP address during the period
dl_site	Total downloads by this IP address during the period
dl_site_ip_agent	Total downloads by this IP/Agent pair in the period
dl_site_session_id	Total downloads with this session ID (Wiley only)
dl_per_day_peak	Peak downloads by this IP address on a single day during the period
total_items_downloaded	Number of items downloaded by this IP during the period
total_items_downloaded_ip_agent	Total number of items downloaded by this IP/Agent pair in the period
total_items_downloaded_session_id	Total number of items downloaded with this session ID (Wiley only)
first_seen	Date of first download by this IP address during the period
last_seen	Date of last download by this IP address during the period
flagged	Boolean representing whether the source data provider labeled this as a robot (1 = true = flagged as robot)
oa_status	Status of the article downloaded. Closed, OA (Hybrid) (the item is OA and the journal is hybrid) or OA (full) (the item is in a fully OA journal) (Wiley only)

[1]Indicators – 1: Agent name. 2: Reverse-lookup. 4: Access/frequency. 8: Other indicator(s)

Table 2. Variables used for determining robots

The first pass through the data involved identifying and labelling self-identifying robots based on the user agent field. The second pass required a reverse DNS lookup for the remaining downloads, and identifying unambiguously human or non-human downloads, based on frequency of downloads, number of items downloaded, and the source of the download, whether internet service provider, server hosting company, university, etc. The third pass on the remaining downloads included querying the database for more information on the session and checking ProjectHoneypot (Unspam Technologies Inc., 2017) for further information on the IP addresses.

Once each item in the dataset is fully labelled as either robot or human, the data will be shared with members of the COUNTER Robots Working Group for peer review to ensure the best possible characterisation of the downloads. At this stage the filters listed in Table 3 will be simulated in the sampled data. A new column will be added for each filter to indicate whether the filter would have labelled the download as robot or human. Each filter and combination of filters can then be categorised as true/false positive or true/false negative, and recall and precision calculated.

Filter	Tests
UA string	Check effectiveness of the COUNTER CoP List of internet robots, crawlers and spiders[1]against manual checking
Rate of requests (single item by a single user[2])	Ascertain best-fit threshold, e.g. within a 24 hour period
Volume of requests (sitewide by a single user[2])	Ascertain best-fit threshold, e.g. within a 24 hour period
Double-clicks (COUNTER CoP 5 7.2)[3]	Determine effectiveness
User agents per IP address	Ascertain best-fit threshold
Request = referrer	Determine effectiveness

[1]https://www.projectcounter.org/code-of-practice/appendices/850-2/

[2]User is defined by best available data: IP address, IP/user agent pair, session ID. This is a proposed adaptation of current COUNTER recommendations (COUNTER, 2017b)

[3]https://www.projectcounter.org/code-of-practice/counter-release-5-draft-code-practice-consultation/

Table 3. Filters to be tested individually and in combination

While the (robot) recall and precision is the basic measure of the filters' effectiveness, of greater concern is how accurate the resulting filtered download/usage statistics will be. In this context, the filtered usage statistics are measured using inverse recall and inverse precision. A low inverse recall indicates that many human downloads were excluded from the usage statistics; a low inverse precision indicates that many robots were included in the usage statistics. We will report on both measures, but will favour inverse precision as the measure of 'accuracy' for two reasons: we assume we will achieve a high inverse recall by default (as the proportion of robots to humans is very high), and because while all the other three measures (recall, precision and inverse recall) are in a way statements about what is excluded, inverse precision is a measure of only what is reported as genuine usage. That having been said, inverse recall will be an indicator of overreach for any particular filter, for example in the case of institutions that access e-resources through a proxy, thereby assigning a single IP address to every member of the institution.

Once the best combination of filters has been determined, the recommendations will be shared with the COUNTER community and feedback sought.

4. Conclusion

Registered vendors undergo independent audit to be considered COUNTER compliant, and for this reason the Code of Practice is generally oriented towards commercial operations. With representatives from PKP, DSpace, EPrints and DigitalCommons on the working group, it is hoped that the recommendations will also have wide adoption within the Open Access community, and that 'COUNTER conformant' usage statistics will become a norm.

As a final note, the filters developed and studied here are not intended to be exhaustive, they are simply a beginning of a set of standards to be built upon that will help in achieving consistent, credible and comparable usage statistics that can be aggregated across many types of scholarly communication platforms.

References

COUNTER, (2017a). About COUNTER. Available at: https://www.projectcounter.org/about/ (accessed 28 April 2017).

COUNTER, (2017b). COUNTER Release 5 Draft Code of Practice, section 'Data Processing'. Available at: https://www.projectcounter.org/code-of-practice-sections/data-processing/ (accessed 28 April 2017).

Greene, J. W. (2016). Web robot detection in scholarly Open Access institutional repositories, *Library Hi Tech,* Vol. 34, No. 3, 500 – 520.

Huntington, P., Nicholas, D. & Jamali, H. R. (2008), Web robot detection in the scholarly information environment, *Journal of Information Science,* 34, 5, 726 – 741.

Information Power Ltd., (2013). IRUS download data: identifying unusual usage. Available at: http://www.irus.mimas.ac.uk/news/IRUS_download_data_Final_report.pdf (accessed 2015-12-11).

IRUS-UK, (2017). Available at: http://www.irus.mimas.ac.uk/ (accessed 28 April 2017).

Sheaffer, R. L., Mendenhall, W. & Ott, R. L. (2006). *Elementary Survey Sampling.* Thomson, London.

Unspam Technologies Inc., (2017). Project Honeypot. Available at: https://www.projecthoneypot.org (accessed 28 April 2017).