

Appendix B

COUNTER/JISC Research Project for Usage Filters: Methodology for Calculating Unique Article Metric

Overview

The purpose of this research study is to investigate the feasibility of new metrics that can serve to eliminate or at least dampen the effect a user interface may have on reporting of usage data. One such metric is the number of Unique Article Requests in a session.

Tests by Elsevier and EBSCO using their COUNTER-compliant usage data have yielded promising results, indicating that it is feasible to apply a filter to the data that will result in a meaningful figure for the number of Unique Article Requests in a session. Furthermore, initial feedback from librarians indicates that they would find this a very useful additional metric.

We would now like to test this data filter with a selected number of vendors, in order to determine whether its widespread implementation will be feasible. You are one of the four vendors we have invited to participate in these tests.

Please bear in mind that this is a genuine research project. The COUNTER Executive Committee will discuss its findings and take them seriously, but is not bound to implement them, as there are other issues that have to be taken into consideration in making upgrades to the Code of Practice. The results that you provide will be a very important factor in our deliberations as to how we can further improve the Code of Practice.

This document describes the methodology that can be used to run tests on this new metric and determine the best way to handle lengthy sessions.

Unique Article Request Count

This is a proposed new metric for COUNTER that would be determined by examining the full text request transaction logs after the existing COUNTER double-click filter has been applied. As its name suggests, this metric is designed to reduce the potential over counting of articles due to the interface effects in which the same article is retrieved in different formats (HTML, PDF, etc.) or for different output options (display, email, print). Please note that the Unique Article Request count would not replace the existing COUNTER usage reports. Journal Report 1 (Total Number of Successful Full-Text Article Requests by Month and Journal). Rather, it is envisaged that this new metric would provide an additional perspective on usage.

Goals of the tests:

- Prove that obtaining the unique article request per session metric is possible

- Calculate the percentage of reduction in total article counts achieved by the unique article count. (For example, if the actual article request count for a given set of sessions was 100 and the total unique articles was 75, then the reduction would be 25%)
- Determine the effect of splitting sessions into time-slices (e.g. to account for long sessions which may represent a series of users using the same library workstation without logging out between).
 - Compare different time-slice lengths and their effect on the reduction percentage.

The prerequisites for the transaction log for calculating this metric are as follows:

1. At minimum all HTML requests and PDF Requests must be logged
2. The logged events include the article's unique identifier (e.g. DOI, PII or other unique ID)
3. The same unique identifier is used for all formats of the same article (e.g. the PDF and HTML of the same article have the same DOI).
4. Transaction log must include an element (or series of elements) that uniquely define a user session (e.g. user session cookie plus Machine Cookie or IP (if Machine Cookie is not available))
5. Transactions within the log for the user session have time-stamps that record the transaction time to the accuracy of at least the second.
6. If multiple servers handle sessions, logs will be consolidated into a single transaction database for the study (all activity captured)
7. If multiple servers handle activities for a single session, the clocks on the servers should be synchronized

Calculating the normal session count

Step 1: Preparing the usage data.

- Remove double-click activities from the transaction logs
- Pre-process usage data as necessary for test. For each transaction the following columns are wanted, at a minimum:
 - IP Address of user
 - CustomerID
 - User Token /SessionID/User Machine Cookie
 - Session time-slice (a sequential number – this will be populated later)
 - Type of request (Search, abstract request, full text request, etc.
 - Format requested (HTML versus PDF)
 - UniqueID of Journal
 - UniqueID of Article (only for item requests)
 - Time-stamp of activity

Step 2: Separate session data into three groups by number of full text requests per session to normalize results with other vendor tests.

- Group 1 contains sessions with zero or one full text requests. The Unique Article processing will have no effect on this group so these sessions will be ignored by the tests.

- Group 2 contains sessions with 2 to 10 full text requests within a session.
- Group 3 contains sessions with greater than 10 requests (this group will most likely represent multiple users sharing a session, or crawlers pulling full text)

Step 3: Calculate the number total number of full text article requests in PFD format in Group 2. The result is “Group 2 Total PDF Full Text Requests”.

Step 4: Calculate the number total number of full text article requests in HTML format in Group 2. The result is “Group 2 Total HTML Full Text Requests”.

Step 5: Using the sessionID, and the unique article ID, determine the “unique” count for articles (without regard for format) in Group 2. The result is “Group 2 Unique Full Text Requests”

Step 6: Report the results and calculate percentage reduction. Following is an active spreadsheet that can be used for this report – substitute your numbers in the “GROUP 2” boxes provided

SUMMARY OF RESULTS FOR GROUP 2	
	Full session
PDF Full Text Requests	50
HTML Full Text Requests	50
Total Full Text Requests	100
Unique Full Text Requests	75
Reduction in Count	25
Percentage Reduction	25.00%

Step 7: Repeat steps 3 through 5 for Group 3 sessions and record results in the following table.

SUMMARY OF RESULTS FOR GROUP 3	
	Full Session
PDF Full Text Requests	50
HTML Full Text Requests	50
Total Full Text Requests	100
Unique Full Text Requests	75
Reduction in Count	25
Percentage Reduction	25.00%

Session Time-slices

The next series of steps will test the effect that time-slicing a session will have on the percentage reduction. The time-slice is a very simple technique of chopping a session into equal time slices of “n” minutes – except for the last time-slice which will be less than or equal to “n” minutes. Tests will be run with slices of 2, 5, 10 and 15 minutes each.

Step 8: Introduce a session time-slice of 2 minutes by reprocessing the usage file and inserting a sequential value in the Session Time Slice field proposed in Step 1. The following pseudo code describes one way of tackling the problem.

```

Sort transactions by sessionID then time
Set time-slice-duration to 2 (varies by test)
For each session
    Obtain timestamp of the first transaction and set to base-time to equal this
    Set time-slice value to 1
    For each full text transaction in the
        If transaction time > base-time + time-slice-duration "minutes"
            Increment time-slice value
            Set Base-time = last transaction time
        End if
    Update Time-Slice-Sequence field in transaction log
    End
End
End

```

Step 9: Using the sessionID, Session Time-Slice and the unique article ID, determine the "unique" count for articles (without regard for format) in Group 2.

Step 10: Report the results in the table below and calculate percentage reduction. Repeat for time-slices of 5, 10 and 15 minutes.

SUMMARY OF RESULTS FOR GROUP 2

	Full session	2 minutes	5 minutes	10 min.	15 min
PDF Full Text Requests	50	50	50	50	50
HTML Full Text Requests	50	50	50	50	50
Total Full Text Requests	100	100	100	100	100
Unique Full Text Requests	75	75	75	75	75
Reduction in Count	25	25	25	25	25
Percentage Reduction	25.00%	25.00%	25.00%	25.00%	25.00%

Step 11: Repeat steps 8 and 9 for Group 3 sessions and record results in the following table.

SUMMARY OF RESULTS FOR GROUP 3

	Full session	2 minutes	5 minutes	10 min.	15 min
PDF Full Text Requests	50	50	50	50	50
HTML Full Text Requests	50	50	50	50	50
Total Full Text Requests	100	100	100	100	100
Unique Full Text Requests	75	75	75	75	75
Reduction in Count	25	25	25	25	25
Percentage Reduction	25.00%	25.00%	25.00%	25.00%	25.00%

Step 12: Analyze results to draw conclusions on the effect of count reductions between Group 2 (normal sessions) and Group 3 (long sessions) as well as the effect of various time-slices on the reduction percentage. Some questions to be answered in the analysis:

If Group 3 results showed a high level of reduction for full session, did one of the time-slices bring the reduction in line with Group 2 results?

Group 2 results are expected to represent true user sessions; therefore, the reduction occurring for the “full session” should represent the “ideal” reduction. Did either the 15, 10 or 5 minute time slice keep the reduction percentage within 4 percentage points of the “ideal” reduction? If so, which ones?

Group 3 represents the long sessions with possibly multiple users therefore the time-slice is needed to prevent over-reduction. What time-slice brings the reduction percentage in line with Group 2 results? What was the effect of the same time-slice on Group 2 results?