



**Usage Research**

## **WHAT TO COUNT & WHAT NOT?**

A White Paper on the Filters to be Applied  
to Raw Usage Data  
Before Usage-Analysis Can Start

### **Version 2.0**

September, 20 2002

**(version 1.0 September 9, 2000)**

**For :** © ScienceDirect ®  
**By:** drs. Marthyn G.M Borghuis  
Senior Manager Usage Research  
ScienceDirect  
Amsterdam

## Introduction

The basic material for producing reports on the usage of electronic products on the Web is the raw usage data either in the format of web-server usage logs or as so-called ‘key-events’ created by the database-server. This applies not only to ScienceDirect®, but also to many of the current publisher web-sites.

With the analysis of usage data a number of critical issues exist, one of which has to do with the filters applied to this raw data. ScienceDirect® wants to achieve the highest possible standard in usage reporting to its customers and has therefore decided to apply new filters to its usage data. How this is accomplished is explained in the White Paper version 2.0

## What is a web-server log ?

A web-server usage log consists of records containing the following information:

- Who requested a URL?
- What URL was requested?
- When was the URL requested?
- How was the request fulfilled?

The key information captured in a usage log file is as follows:

### Who?

The IP-address of the end-user's computer, the username (when a personal login exists), and session and/or user cookies

### What?

URL of the page requested

### When?

Date/Time stamp (*dd/mm/yyyy;hh:mm:ss*) and the time zone (indication referring to the time at the end-user's location to GMT)

### How?

Web browser used including version number (e.g., Netscape 4.6, Microsoft Internet Explorer 5.0, etc.) and the operating system (e.g., Windows 98, Mac, etc.)

Additional usage log information\* includes the following:

- Return code (a code indicating the status of a request, e.g., successful, failed, refused, etc.)
- Total number of bytes transferred with a request.
- Referrer or previous URL (page from which a request or mouse-click was made).

\* With the further development of browser/web-server software, the number of items that can be logged will increase.

## Objective

An end-user, logged on to an Internet site, generates at least one record in the web-server log file with each click of their mouse. A *raw* web-server log file, however, contains a multitude of records that are not related to an end-user's requests. In order to effectively report the intentional usage of an end-user, it is essential to exclude the unrelated log records from the web-server log file prior to performing the usage reporting analysis. Filtering the records in the web-server log file will reduce the records reported to only those records which correspond to what the end-user actually requests to see and their entitlements.

For reliable reporting it is key to include only the log records that can be identified as intentional usage of an end-user. It is the objective of this paper to define both, the rules governing the implementation of filters to reduce the number of records analyzed and the filters used in this process.

## The Basics

One mouse-click by the end-user may produce multiple log-records.

### Example 1:

An end-user clicks on an http-link to retrieve a page. This page contains embedded images (GIF, JPG or JPEG files) or could contain other objects that do not change with the content of the page. Records are generated in the log file for the page requested by the end-user and for each of the images/objects *requested* by the page. These latter records are called "server generated logs". These records should be removed.

### Example 2:

An end-user clicks (or requests) a page which he/she not authorized to access. The log file generates a record of the user's request with a return code '401' (unauthorized) or '403' (forbidden). The end-user does not receive the requested page; therefore, these records should be removed.

**Note:** Retaining these records outside of the usage reports could be needed for special analysis on how many unauthorized users request access.

It is clear that many log records should not be retained. Approximately 50% of a full web-server log file should be filtered.

A basic filter will remove:

- All log records generated when a page requests the embedded image files, e.g., .GIF, .JPG, .JPEG, frames and style-sheets (.CSS) etc. for the page. (See Example 1 above).
- All log records containing a return code other than "200", (successful request) or "301" (redirect) or "304" cached page. (See Example 2 above.)

After this basic filtering, the web-server log file contains only successful and intentional end-user requests.

## One Step Further...

### Multiple successful requests for PDF files created by the browser software

In a close look at the successful requests which are recorded in the web-server log file, it has been found that Microsoft® Internet Explorer - Versions 4.x and 5.x create multiple successful request records in the log when an end-user requests a PDF (Portable Document Format) file. As most publishers' articles on a Web site are in PDF-format, this causes the appearance of an exaggerated number of articles in PDF-format requested.

There may be other browsers, which are creating the same (or creating even more) multiple successful requests records. Because new browser software and/or new versions of existing software are released frequently, it will be very difficult to identify all occurrences of this misrepresentation in the web-server log file.

### Unintended successful requests created by an end-user

The multiple successful requests for PDF files created by the browser(s) are the affect of the software and, thus, not related to end-user behavior. End-users, however, do other things at their desk-top when they are browsing a Web site which causes multiple entries in the web-server log file.

The most common examples of unintentional requests include the following:

- An end-user may accidentally double-click on an http-link where a single click would suffice. A double-click is normally used in Microsoft Office to open a program or a file. A double-click on the Web results in two (2) successful requests recorded in the web-server log file.
- An end-user may press the **Back** button on their browser to go to the previous page within a PDF file. This action results in an additional successful request recorded when they go back to the PDF file.
- An end-user may press the **Refresh** (or **Reload**) button on the browser. Again, an additional successful request is recorded for the same PDF file to reload.

As a result, within a limited time period, multiple entries occur in the web-server log file by one end-user (or IP-number) for one Web page (or URL).

## Developing Decision Rules

One approach in the mapping of the problem and the creating of the decision rules is to document all the different occurrences of multiple entries in the web-server log file. This is a cumbersome exercise, not only because new software is released so frequently, but also, because the occurrences involving the retrieval of PDF files (see the *One Step Further...* section of this document) are often not identifiable because of missing browser information in the log file.

Moreover, the unintentional requests (also see the *One Step Further...* section of this document) are all based on the interpretation of end-user behavior. End-user behavior will always be open for differences in opinion.

For these reasons another approach was taken. By monitoring what happens when multiple entries in the web-server log file are removed based on an increasing time-window, the following determinations were made.

## Testing the Decision Rules

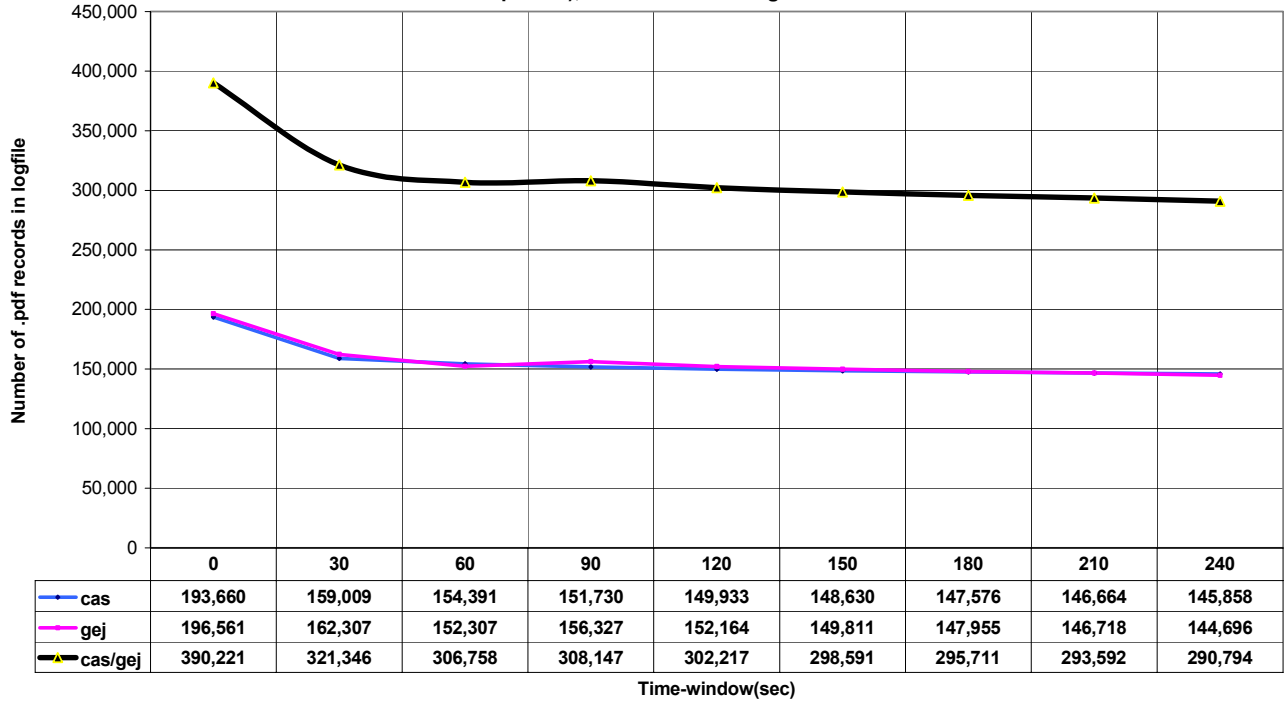
### Full Text Servers for Elsevier Science Generic Electronic Journals

A test was carried out for the two (2) servers holding the full-text articles for Generic Electronic Journals (GEJ's). These servers are called: "CAS" and "Deja-Vu". For the analysis, a sample of almost 400,000 web-server log file records for full-text articles in PDF-format was used as generated on two (2) different servers, ww1.elsevier.nl server and ww3.elsevier.nl server. These servers hold the full text of Generic Electronic Journals

For testing purposes, an identical record is defined as a record containing an identical instance of the same information which includes all of the following:

- one IP number requests
- one URL
- a defined time-window

**Test 1 for PDF files of GEJ articles :**  
**Change in absolute figures when removing identical requests for articles in pdf (i.e. same user requests same pdf-file), within an increasing time-window.**



**Test 1 Results**

When the time-window is set at **60 seconds**,

$$\begin{array}{r}
 390,221 \text{ total number of requests found} \\
 - 83,493 \text{ requests removed} \\
 \hline
 306,728 \text{ end-user requests (or a reduction of } \mathbf{21.4\%})
 \end{array}$$

Requests removed include:

- All identical requests caused by double-clicking the http-links
- All requests generated when end-users request a PDF file via Microsoft IE 4.x and 5.x.
- All identical requests caused by pressing the Back button and the Refresh button while the server is still transferring a page (e.g., downloading a PDF file, etc.).

When the time-window is set at **90 seconds**,

390,221 total number of requests found  
- 85,915 requests removed  
 304,306 end-user requests (or a reduction of **22.0%**)

Requests removed include:

- All identical requests caused by double-clicking the http-links
- All requests generated when end-users request a PDF file via Microsoft IE 4.x and 5.x.
- All identical requests caused by pressing the Back button and the Refresh button while the server is still transferring a page (e.g., downloading a PDF file, etc.).

**Test 1 - Conclusion**

Larger time-windows generate a greater but only marginal reduction. Obviously, the larger the time-window used, the higher chance that identical requests are the results of an end-user's wish to open again (or re-visit) the same file or page.

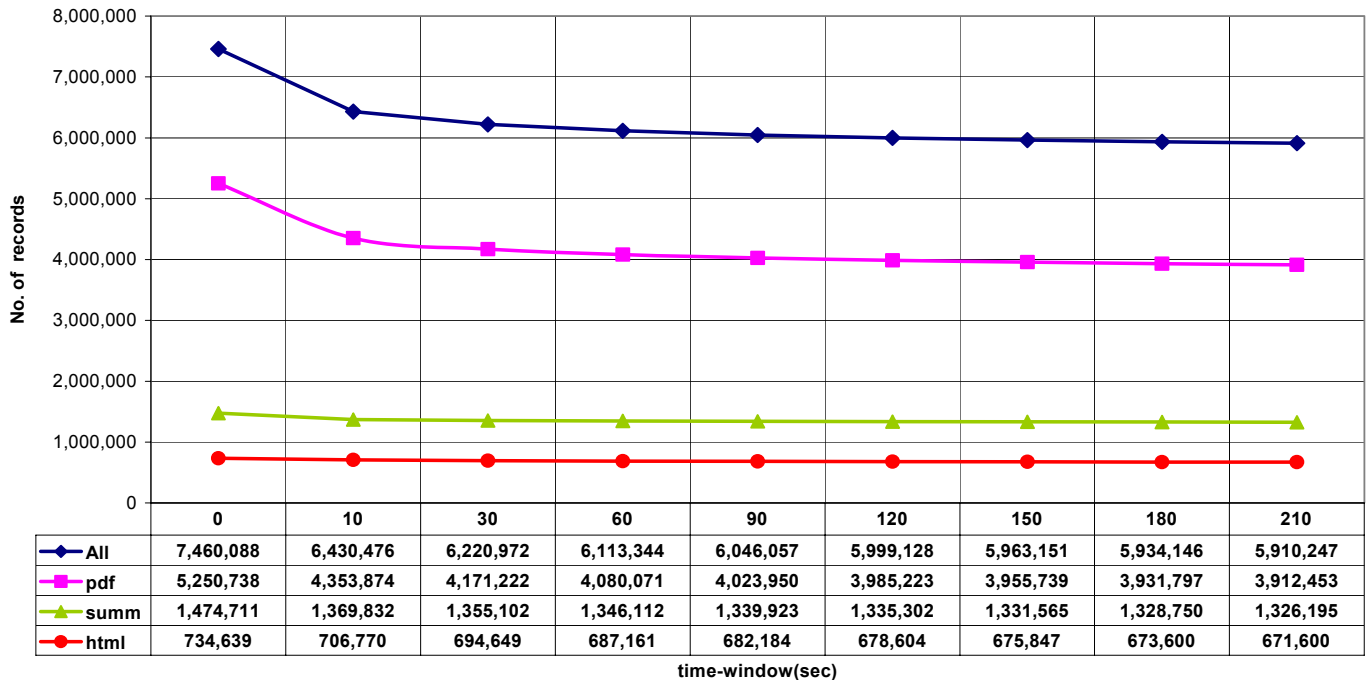
The time-window to use for filtering identical request records from the web-server log file should be **60 seconds**

**Applying the Decision Rules**

**ScienceDirect Web-Server logs**

The same analysis was carried out for a sample the ScienceDirect web-server log files for April 2002. These logs refer to usage of full-text articles in the two (2) formats: PDF, HTML and the usage of the ScienceDirect special format : Summary Plus. Just over 7.4 million records were included in the analysis.

**Test 2: ScienceDirect**  
 % change in absolute figures when removing identical requests within increasing time-delays



## Test 2 Results

### *HTML usage*

Looking at the reduction of HTML and Summary (which is also in html format) log records, an almost flat line appears in the graph, indicating the removal of double-clicks and the identical requests caused by MSIE 4.x and 5.x. This confirms the earlier statements made. These multiple requests will occur within a **10 seconds time window**. This **10 seconds time window** is thus applied to remove all multiple requests

### *PDF usage*

On the level of requests for PDF, the graph shows a nearly identical pattern as compared to the results under test 1, only the time elapse between two identical requests has changed dramatically and is now on a **30 seconds time window** in stead of 65 seconds as was the result of the analysis in the year 2000.

- In the 10 seconds time-window, the reduction is  $5,250,738 - 4,353,874 = 17.1\%$
- In the 30 seconds time-window, the reduction is **20.6 %** and stays almost flat at higher time-intervals

## Overall Conclusions

This White Paper has defined the decision-rules for filtering those records from the raw web-server logs that do not refer to real and intentional end-user behavior. After applying the filters, the web-server log files should contain only successful and intentional requests.

This filtering should remove:

1. All images and other server generated log records.
2. All log records with a return code other than 200, 301 and 304 for certain websites.

Detailed analysis of web-server logs from two (2) different servers pointed out that the peak in the occurrence of identical requests for pdf's (same IP/user requesting same URL) occurs at the 30 seconds time-interval between 2 identical requests. After 30 seconds only a very limited number of double clicks will be removed. Identical requests for all html pages occurred within 10 seconds. Therefore, in addition to the 2 filters described above under 1. and 2., a third and fourth filter should be applied to remove:

3. All identical requests occurring within the 10 seconds time-interval.
4. All identical requests occurring in the 30 seconds time-interval for requests for PDF files only.

All four filters are applied to all web-server logs and database 'key-events' as follows:

- ✓ For products running from the ScienceDirect platform, filtering was implemented starting October 2000. From Aug 2002 on the 30 seconds time window for removing multiple requests for PDF files was implemented.

*In case you need any further information please contact the Author at the following e-mail address:  
M.Borghuis@Elsevier.com*

**Amsterdam, September 20, 2002**