

The Future of Traffic Analysis on the Web

Philip Hunter, Information Officer at UKOLN.

Abstract: Web statistics have been a bone of contention for some considerable time. Part of the reason for this is that the analysis of the statistics now serves a quite different range of functions from those originally envisaged.

What is Web Traffic?

The logging of information about visitors to Websites began as a simple extension of the information kept by server administrators long before the Web came into being. In those days the information about files served, peak times for traffic., etc were collected in order to provide detail about the level of service. This could be used by the system administrator to anticipate necessary upgrades to the service, to spot bottlenecks in processing, identify unauthorised use of the facilities, etc. Analysis of this data could be presented in simple graphical form to the managers of the service in hourly, daily, or weekly breakdowns, illustrating aspects of usage.¹

Then the Web arrived. At first there wasn't a problem, since most early sites were academic-related, and inline images were not an option. A hit to a page was exactly that, and the statistics told you pretty much what you wanted to know. Was anyone looking at your page? Which domains did they arrive from? Which is the busiest or quietest time? Are users attempting to find pages which have since been moved? Etc.

Once inline graphics became possible in mid '93, statistics for the Web became a more complicated issue: a hit to a page with ten or so inline graphics (graphical bullet points, section dividers, headings, banners, etc) would result in ten further requests. In other words the statistics began to reflect the complexity of documents served rather than the usage of files. It was still possible to work out what was going on, but no longer straightforward.

The arrival of large-scale commercial interest in the Web in 1995 changed everything. The Web was no longer an exclusive playground for IT people and academics, and the newcomers needed statistics for their own purposes. Marketing departments needed stats to show that the investment in a Web presence was generating some kind of interest, if not direct evidence of business done, and to establish relative success compared to competing businesses.

For a while hitcounts were bandied about as if there were no problems associated with them. However commercial companies depend for their long-term success on being able to spot dud information. With the increase in the use of proxy servers and caching services, it became impossible to treat a raw hitcount by itself as reliable evidence of usage.

Making Web Statistics Reliable

Where we are now is that all access statistics are treated as suspect unless the parameters within which the statistics have been collected and processed are explained in detail. This means that it should be possible in theory for a third party to understand the significance of the data, and to reprocess it in a format which makes it possible to compare it with data from another site. In other words, in the early days of the Web, it was assumed that Web statistics data were straightforwardly intelligible and unambiguous in meaning, and therefore overtly comparable with each other. Now the assumption must be that both data collected in Web logs, and data presented in a report are *not* straightforwardly intelligible and unambiguous in meaning, and not comparable until the parameters of collection and processing are both stated and intelligible.²

There is as yet no standard by which these statistical analyses are prepared, though there are discussions under way about putting such standards in place.³ An example of the sort of standards we might require in analysis would be (for example) presentation of only those visitor sessions not using the same IP address within thirty-minutes. Obviously this isn't the only way in which visitor sessions might be analysed and presented. However, like the collection of distribution figures for magazines and

newspapers, and the assessment of the audience for television programmes, it is more important that a method for the collection of figures is agreed on, than that the method be shown to be without weakness. So, as in the case of the circulation of these conventional publications, we need a reliable way of *auditing* Web traffic.

Making more of Traffic Analysis

Before discussing the various ways in which useful information can be extracted from Web statistics, and discussing the additional elements which have to be recorded to give information about transactions useful both to academia and to commerce, it is worth looking back at the development of traffic analysis in WWII, where it is possible to see the antecedents of some of the complexities involved in Web traffic analysis.

Both Gordon Welchman, formerly a Cambridge Mathematician and later a cryptographer at Bletchley Park, and Josh Cooper, head of the Air section, made some interesting observations (long after the event) about the situation before the Enigma codes were being broken on a regular basis. The radio stations around the UK which monitored the radio traffic from German sources were part of what was known as the 'Y Service' (one each for air, army and navy). Cooper said that GC&CS (the Government Code and Cypher School, the antecedent to GCHQ) 'had always tended to take too little interest in the radio by which they lived'. And the Y services were correspondingly dismissive of the work of the codebreakers. The Y services believed they produced sufficient intelligence simply by analysing the activities of the radio networks they were monitoring. In fact the RAF site was completely ignoring the Enigma traffic. When Cooper suggested that an outstation should begin taking Enigma he was told by the head of the RAF Y Service that: 'My Y Service exists to produce intelligence, not to provide stuff for people at Bletchley to fool about with'.

One of Welchman's first jobs was to study station call signs and the preambles to messages in the hope that these might include information that would help to solve the Enigma problem. He later wrote:

Previously, I suppose I had absorbed the common view that cryptanalysis was a matter of dealing with individual messages, of solving intricate puzzles, and of working in a secluded backroom with little contact with the outside world. As I studied that first collection of decodes, however, I began to see, somewhat dimly, that I was involved in something very different. We were dealing with an entire communications system that would serve the needs of the German ground and air forces. These callsigns came alive as representing elements of these forces, whose commanders at various echelons would have to send messages to each other. The use of different keys for different purposes suggested different command structures for the various aspects of military operations.⁴

There is a parallel here with the analysis of Web statistics: the collection of basic intelligence about Web usage resembles the collection of callsigns and frequency data by the wartime 'Y' Services. The actual transactions - which contain information about what the users are doing on the site, and what their interests and intentions are - parallels the information encrypted in the Enigma codes (containing detailed user information). Detail of this kind is currently scarcely catered for in the Web logs and the analytical tools built for them. We are not in the position where we can gain a full picture of what is going on.

This means that current analyses of Web traffic are lopsided, and do not give anything like the insight into the usage of Websites which they might. That is to say that the data analysis packages generally available represent the high end of what is possible with currently available log file information. But, since the logfiles were originally designed as the logical development of what pre-Web service administrators required - information about bandwidth usage, file usage, user logins, user frequency, etc - they do not contain the kind of information required in order to provide useful profiles into the sort of e-commerce activity which represents the future of many species of Web services.

The principal requirement (whether they understand this or not) for successful companies and institutions employing Websites and services for e-commerce purposes is to have the capacity to create a total picture of the usage of their services. As Gordon Welchman realised that he was dealing with 'a total communications system', so the analysts of Web traffic require information about requests and transactions *which add up to a total picture of usage*. Analysis of the data need not of

course involve all available information, but the whole range of information should be available on demand in order to maximise the use of any particular data set within the whole.

What this means is that instead of trying to extrapolate information from a narrow and incomplete dataset, which is the current situation, the opportunity would exist for the significance of any particular piece of data to be interpreted in the context of a complete dataset of server transactions, including both basic traffic data, and detailed information about usage and transactions made via the service. It would be possible to build up a complete picture of site usage, and to prepare reports on any aspect of the patterns of transaction.

The Shape of Future Web Traffic Analysis

There are currently two main kinds of institution which need to develop richer patterns of data collection and analysis, in order to get the best out of the Web as a transactive medium. The institutions are: academic organisations, and commercial companies. A third type is of course government institutions. Since not every government transaction involves financial operations, it is probably best to view government on the Web as a hybrid of the technical development of the other two. Here we are going to look briefly at some of the issues for the future of both academia and commerce on the Web.

Many universities are now using the Web as a way of delivering content to students locally and remotely. Educational content is necessarily assessable in order that students can be graded throughout courses, and this means that interactive content is of much more use to the institutions than 'flat' content. The university needs to record the interactions, both in order to assess the progress of the students at or near the event, and also for the use of the university's management and information system (MIS). Standardised formats for recording these details are also desirable for the processing of this information. This does not mean that educational courses will be put into straightjackets when it comes to the designing and the implementation of assessable courses, only that the standardised formats and systems used have to be

sufficiently flexible to accommodate the academic and institutional purposes of the educational materials (for example XML has provided a large degree of flexibility while still offering interoperability between systems). Much of this area falls under the relatively new area of Instructional Management Systems (IMS).⁵ The IMS and the existing Management Information Systems (MIS) currently running in most universities have to be able to talk to each other, and it should be possible for the expanded range of logged information passing around the institution to be subjected to the same kind of automated analysis as are conventional Web stats. The general reports will be larger, since there is more information in the logs, but the larger amount of information is most important for the granularity of view which can be generated: there will be more depth of detail for mining.

For commercial companies, the principal data to be mapped by augmented Web logs is customer transactions. This information is already collected, one way or another, however it is not always recorded in a way which can be easily analysed. The information is not represented in the conventional Web log, and cannot therefore be processed by standard analytical tools. As things stand, the e-commerce information may well be split up among several aspects of the service. What is required is that the data about wish lists, shopping cart orders, actual orders completed, transaction values, customer information currently stored in cookies on the users machine, etc., should be stored in a standardised format compatible with the existing Web logs. Every aspect of the customer's journey towards a successful transaction should be available for scrutiny, since analysing the behaviour of the customer is the best way of understanding the strengths and weaknesses of the Website as a vehicle for commercial transactions.⁶

In both cases the academic institution and the commercial company require to collect information of sufficient richness to reconstruct and analyse all assessable processes and transactions. This is the bottom line in the development of accountable electronic services.

One of the developments appearing over the horizon which may complicate the issue is the newish notion of 'Web Services'. These services are likely to produce a centrifugal tendency in data storage - not all related transactions will happen in the

same location. The total data picture will be distributed until collected, and therefore not a whole picture at all, until invoked. Current analysis packages are (naturally) not set up to deal with information held in this way. Do we re-write all the software tools? Perhaps one of the first of the Web services which should become standard is one which - by default - harvests all relevant data from all parties to transactions. Otherwise what we gain in distributed services might well be offset by an inability to gain an accurate impression of the big picture.⁷

¹ An example of this kind of analysis can be seen at:

<http://www.bath.ac.uk/About/Usage/Departments/Library/Monthly/2001-08.html>.

² Electronic Journals have strong commercial reasons for requiring accurate data about usage of their articles. There is an interesting White Paper by Judy Luther on this issue at: <http://www.clir.org/pubs/reports/pub94/contents.html> (CLIR Report, October 2000). Also available as a reprint at: <http://www.press.umich.edu/jep/06-03/luther.html> (The Journal of Electronic Publishing, March, 2001, Volume 6, Issue 3 ISSN 1080-2711)

³ The UK PALS group was set up to explore issues of interest to publishers and libraries ('Publisher and Libraries Solutions') <http://www.library.yale.edu/~llicense/ListArchives/0104/msg00016.html>. A related group is exploring usage statistics - details can be found on the JISC site at: http://www.jisc.ac.uk/curriss/collab/c6_pub/#uswg.

⁴ Smith, Michael, *Station X: The Codebreakers of Bletchley Park*, Channel 4 Books, 1998. From Chapter 3 'The First Break' p24-5.

⁵ Gardner, Tracy, *A Developers Perspective on IMS*: In Ariadne issue 20, June 1999. <http://www.ariadne.ac.uk/issue20/ims/intro.html>.

⁶ Quantified Systems, Inc., distributes plugins which make it possible to include data about transactions in a separate log which can be analysed together with the standard traffic log. A white paper giving details of how this is done is available at: <http://www.urchin.com/support/whitepapers.html>.

⁷ Gardner, Tracy, *An Introduction to Web Services*: In Ariadne issue 29, September 2001 <http://www.ariadne.ac.uk/issue29/gardner/>.