



"turning data into dollars"

Tom's Ten Data Tips – May 2007

Web Usage Analysis

Web Usage Analysis is one of the frontiers in data analysis. Because every mouse click gets recorded, every page viewed, but also when, you can closely watch in minute detail every step a person takes on the web. At the moment, we are grossly lacking the conceptual models to fully exploit the richness these data might offer. Not only which pages are viewed in conjunction, but also the chosen navigation paths can offer tremendous insight.

In a recent study, CMS Watch (an independent analyst firm) found that increasingly, e-mail campaigns, keyword bid marketing, and customer segmentation are integrating. As a result, web usage analysis is becoming more of a mainstream concern for analytically focused organizations.

1. The Web Is A Graph

The web is, mathematically speaking, a huge graph, where every node is a web page. Clickstream data record each and every link between nodes, and also *when* it was requested. This graph model has proportions thus far unknown. We have only scratched the surface when it comes to fully exploiting the data riches this has to offer.

The internet is by far the largest distributed system we have ever known. As a result, we need to adapt our analytical techniques to this "new" reality, and develop algorithms with sufficient scalability to handle these dimensions. In coming years we are bound to see breakthrough developments in technique and theory. In Meteorology, for instance, it took decades before the power of super computers could productively be used to improve on the weather forecast of 3-5 days out. Organizations are only beginning to populate webhouses with the back-end data of their sites.

2. Web Phenomena Reproduce At Multiple Scales

Through research a fascinating theme has surfaced: many internet phenomena occur at *both* the micro and macro level. What this implies is that by "merely" observing the variable distributions of parts

of a site, we can draw meaningful conclusions about other parts of the web *we haven't even seen*.

There appear to be grossly logarithmic variable distributions, that have been reproduced across samples from thousands, to hundreds of thousands, and many millions of page views. The theoretic implication is fascinating, and the practical implication very powerful. From observing a 'mere' variable distribution, we can make inferences as to how it relates (is connected) to the rest of the site, and what expected variable distributions elsewhere in the site should look like.

3. The Web Is The Most Powerful Testing Playground Ever

There is no place like the web for testing. Whether it is ad creatives, promotions, premiums, or the attraction of various product descriptions, immediate click through rates provide outstanding substance for testing. All it takes is a design based on sound methodology, and a clear business goal one is trying to optimize against (see also tip# 4).

It is easy to post a contact strategy and split test variations "live" on real customers, getting real offers. All issues regarding research validity dissolve, and moreover, due to the real-time nature of the web you get instant results!

4. You Can Only Optimize If You Know What Your Goal Is

The power of web usage analysis should lie not only in describing what happened in the *past*, but rather in adapting the site for further improvement in the *future*. For instance, by segmenting users on the basis of pages viewed, one can arrive at a segmentation of visitor needs.

The next step might be to structure the home page accordingly, if helping users satisfy their information needs as quickly as possible is your goal. This is valid for a utility/transaction site. Portals, on the other hand, aim to keep their visitors on the site as long as possible, in order to expose them to as many banners or sponsored links as possible. This all depends on the goal you are pursuing with your site. This demonstrates that the site's goal needs to be made explicit in order to determine what objective you want to optimize.

5. Behavioral Targeting Has The Future

When visitors are served sponsored links or targeted banners, this is still most often done on the basis of search keywords, and pages requested. Search marketing has had a profound impact on the media/advertising business. This is because someone who is actively searching has displayed *some* interest, and the offerings are matched to that. Still, matching to a single search does not have much depth. However, more profound insight in what someone has been looking for, is not very far away: his web usage behavior prior and following the search.

When web analysts speak of "behavioral targeting", this is what they refer to. Mind you, in many cases, the "behavior" being targeted need not be very elaborate, as the term has become the fashionable concept to promote.

6. Page Tagging And Server Logs Both Have Merit

There are two ways to collect clickstream data: through page tagging and by offloading the server logs. Server logs are recorded when a page and its associated objects (images, Flash, etc.) are collected. Page tagging is when each and every page gets a piece of javascript that gets triggered when the page is requested. The javascript then creates a dedicated log on an ASP server.

Both have advantages and disadvantages. Advantages of an ASP solution are:

- Data is "ready" to analyze
- Client information is captured, like which browser they are using, screen resolution settings, time zone, etc. (to collate this from server logs can be erratic and unreliable)
- You can capture outbound links that have been followed

Advantages of Server logs are:

- Data is "free", versus possibly a monthly fee and vendor lock-in with the ASP provider
- You get spider/robot information, which is not available off an ASP solution
- You capture server errors

7. Track As Few KPI 's As Possible

Because there are potentially so many metrics one can calculate, there is a serious risk of "analysis paralysis." When asked what they'd like to know, collective business partners can sometimes dream up an

overwhelming number of metrics to track. Displaying more numbers, does not necessarily lead to more insight, though.

One needs to do an in-depth analysis to find out *which* parameters truly drive the business forward, and explain *why* these are Key Performance Indicators. All other metrics can still be reported on an ad hoc basis.

8. Attempting To Prepare Clickstream Data "Exactly Right" Is Futile

Clickstream data can be rather cumbersome to prepare, especially server log data (see also tip# 6). Since there is a premium on good results fast, rather than perfect results later, a sensible trade off is needed here. Also, keep in mind why data is being prepared, and what it is you are trying to measure. This should aid in making an educated guess when "good" is good enough.

Trying to make totals match from separate web analytics systems is troublesome. Yet the business may wonder why there can be differences in a count as simple as the number of visitors? Clicks recorded on different systems simply don't mean the same thing. Explaining *why* is an important and challenging task for analysts.

9. Integrating Web Analytics With Internet marketing Is Key

It is not enough to just "know" what is going on at your site. Increasingly, internet customers expect you to act on this knowledge, preferably in real time.

Merging email marketing, site personalization, Search Engine Marketing (bidding for *the right* keywords to drive *the right* traffic), Search Engine Optimization, making targeted offers (behavioral targeting), and dynamically adapting content is not an easy task. And not all web solutions are equally well equipped for these challenges. For example, email marketing has been demonstrated to work best when the email subjects are associated with pages/products a prospect looked at, but didn't buy yet. Few companies are currently at this level, yet.

10. Exploit The Near Real-Time Nature Of The Web

Because of the immediacy of conclusions that you may derive from your site, feedback loops from web analytics are exceptionally short

(see also tip# 3). Also, the sheer numbers of any phenomenon we observe on the web make statistical inference in many cases a non-issue: either an effect is there or it's not. The question of statistical significance rarely occurs on a site with even modest volumes of traffic.

What we have failed to observe so far, is organizations taking advantage of these real time facilities. For example, if commercials are running at prime time, even a small variation in the execution will lead to a differential in the number of people who will subsequently visit the site. And you know the "winner" immediately after the last commercial has run! Such is the nature of synergy that web usage analysis can offer to marketing tactics in diverse areas. These opportunities are currently woefully underused.