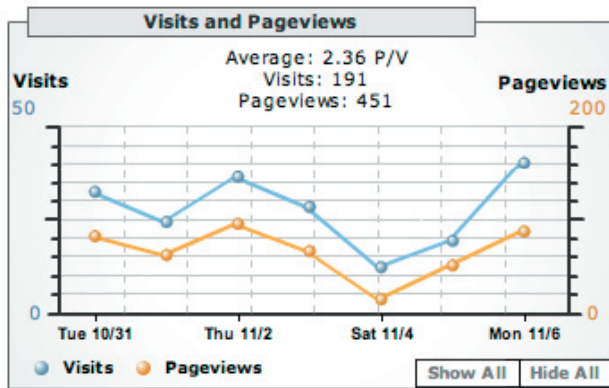


Handout 6: Statistics and Analysis Basics



The first question many people hear after launching a web site is, “How many hits are you getting?” Before you politely tell the questioner that “hits” are an incredibly useless measure of real site usage (and you should, by all means, tell them that), you might want to understand exactly what hits are and how they’re calculated.

A “hit” is generated whenever someone accesses a file on your web site. Pretty straightforward, right? Well, actually, it isn’t straightforward at all. Because a hit is generated for every single file on a site, including all the files within a page – e.g. every single image, no matter how small – the number of hits your site receives is directly impacted by how many files you use to create a page. If you’ve got a page with a lot of text and a single product image, that page will garner two hits when it’s accessed – one for the page itself and one for the image.

On the other hand, a very similar page where all the links and navigation are images – not an uncommon scenario at all – might garner 50 hits every time it’s accessed. That’s because every image on that page, no matter how small or how many times it’s been accessed on other pages, will generate a hit. It’s quite easy to create two pages that look very similar but one will generate ten or twenty times as many hits as the other. Hits, then, are a pretty useless measure of how your site is doing, because one site could see twenty times as many hits as another and still be serving up the same number of real pages.

Point of View

Real pages are, in fact, what we’re really looking for when we examine a site’s raw traffic numbers. After

all, the goal of a web site is to get people looking at your products and services. Every time they look at a page you’re getting product or company information in front of them, and the more they look at, the more likely they are to buy something. The statistic that really matters in this case is called a “page view.”

A page view disregards all those different files that generate hits and simply tells you how often a user viewed a page on your site. (Don’t confuse “page view” with “impressions”; impression is a term used by advertisers to measure the number of times an advertisement is viewed. It’s almost the same thing, but an impression is really a subset of a page view.) For most web sites, page views are the most reliable indicator of actual site traffic.

Even page views, however, need to be viewed in the context of the structure of your site. Ultimately you want people looking at product pages and other areas of the site that translate into sales. If you’re getting 100,000 page views on your home page and only a few anywhere else, you’ve got a problem – that means visitors are abandoning your site after their first look (abandonment is actually an important – and very advanced – statistic that we’ll examine in a future column).

Page views, then, ideally should be somewhat balanced across the site. Most statistics reports return page views by page or by section of the site (or both). If you have a popular product that is attracting a lot of visitors, it’s normal to see that product page or section getting many more page views than the rest of the site. But you should also see page views elsewhere – in your shopping cart, in your ordering information page, or on your “Contact Us” page – if that popularity is actually translating into real orders. Truly “balanced” page views across all the areas of a site are fairly rare, but be wary of too big a spike in one area that isn’t backed up by others – it may indicate that users are checking something out and then bailing out.

Page views may be the single most important statistic for determining how much traffic your site is receiving, but it’s just a part of the larger picture. On a basic level, page views answer the What and When of web statistics: What your users are looking at and when they looked at it. It also provides you

with trend information for site traffic by answering the big question of How Many? The Who and Where are quite a bit trickier, especially if you have a website that doesn't require users to register in order to make purchases.

Getting Real

Page views don't give you a complete picture of your site's success. If all those page views are coming from your web designer admiring his own work or from search engines and "robots" indexing your site, they're not really translating into customer interaction with your products. To view the whole picture of your site's success, you need to find and measure the "who" and "where" of all those page views – specifically, who is visiting the site and where they're coming from.

There are essentially two radically different ways of measuring who and where; the method you choose may already be a foregone conclusion depending on how your site was built. You can always add statistical analysis tools to just about any website, but the most advanced technique for identifying who and where is built in to many e-commerce packages: identifying users through registration.

Welcome Back

If you know what each of your users is buying – and they know what they bought – many common customer service tasks can be partially or fully automated. Not surprisingly, user registration pays big dividends in understanding site statistics as well. By asking users to register with your site in order to use it, you create an incredibly valuable database of user behavior – a detailed "who" and "where" to go with the "what" and "when" of page views.

User registration can be simple or complex depending on the technology you or your service provider choose. Some systems track the users' "critical" activity – such as making a purchase or a support request – and little else. More advanced systems actually paint the whole picture of user activity: When they last logged in, what pages they visited, when they "bailed out" or left the site.

If you don't know what kind of built-in tracking your site supports, ask your service provider or web developer what they've got to offer. If your ser-

vice provider or developer supports it, you should require or at least suggest some method of user registration on your site, even if it means offering perquisites – such as some special feature or one-time discount – to users who register. The value of a known user on your site is many times that of an unknown, anonymous user.

Crunching More Numbers

User registration, however, is not for everyone – you may not be able to force customers who are simply browsing to register, and some customers simply refuse to register for anything or, even worse, provide false information. That's where statistical analysis packages come in. Ideally, user registration is combined with statistical analysis to provide a detailed picture of user behavior. But it's not uncommon to use more advanced web statistic software to track many user activities as well.

Web statistics software runs the gamut from free to extremely expensive, and the maxim that you get what you pay for generally holds true with this kind of software. However, don't be surprised if your service provider has statistical tools available that you've never even looked at – you might be unaware of everything that's under the hood of your web site, so don't hesitate to ask what's available.

Statistics packages can do two basic things: Analyze the log data every web server creates and create more detailed statistical information on their own. Some packages do both, but we'll tackle each function separately for the sake of simplicity. To understand how statistics packages work, you have to go back to that marginally useful piece of data that I showed so much disrespect for in the last column: the lowly hit. Our poor friend the hit – while pretty useless on its own – is actually the grain of sand that creates the beaches and dunes of complex statistics.

Hits are created when a user accesses a file on your web server; the server records all these hits in "log files," which are typically huge, detailed diaries of every file accessed on the site. The log records the time, the file accessed and the IP (Internet Protocol) address of the party on the other end that made the request. Nearly all the statistics that we've talked about so far – other than user-specific stats from registration features – are calculated by analyzing this log file. Page views, hits, time on site, and even

originating location can be determined with log file analysis.

Log analysis, then, can give you much of the “what,” “when” and “where,” although the last one is a little tricky. Log files can tell you where someone came from – that is, their IP address – but IP addresses aren’t always a reliable indicator of where someone actually is. If you know your or your customers’ typical IP ranges, you’ll be able to get some significant use from IP addresses in that they will confirm things you already suspect. But since IP addresses can be faked – “spoofed” is the technical term – and often are faked solely as a security measure, it’s not uncommon to see a good chunk of originating IP address information that is completely useless for identifying where someone came from.

High-end Analysis

Log analysis is probably the most pervasive method of measuring a site’s success, and your service provider should supply you with some form of log analysis at no cost, although they might offer upgrades to higher-end packages if you want more detail (those packages cost them money as well, so don’t expect them for free). But if number-crunching is critical to your business, consider the second option – packages that create their own statistics.

These high-end packages are available both as standalone software and as services that you subscribe to. They typically involve a little bit of customization on your web site – you add a snippet of code to every page or page template, and the system generates a detailed entry of what happened when that page was accessed. By using their own code, these systems are able to track in incredible (and sometimes a little bit frightening) detail what your users are doing as they visit your site. These packages usually don’t come cheap, but if knowledge of your users’ habits and activities is vital to your web strategy, these packages are the holy grail.

Sidebar: Google Analytics

Once again, Google has come along and bought up a prominent software company and re-tooled their product, then turned around and offered it for free. The product in question is an analytics tool called Urchin, which is now called Google Analytics.

Google bought Urchin because they wanted to provide their massive advertising engine with a simple way to track the success of campaigns. By offering an analytics package for free, Google enables any advertiser to track site statistics, campaigns and conversions with unparalleled detail and accuracy.

However, because Google Analytics is free, you can use it to your advantage even if you’re not advertising online. Google Analytics supports highly detailed tracking of visitors and includes sophisticated statistical analysis such as first-time vs. repeat visitors, trends over time, and geographical origins.

To get Google Analytics, just go to <http://www.google.com/analytics> and apply. You’ll need to insert a very simple piece of code in every page of your website (ideally, you’ll want to do this in a header, footer, or “wrapper” file so you only have to do it once) and Google will start tracking for you. It’s that simple, and it’s free.