

## **Using Web Metrics to Estimate Impact: IV – The “Path” to Understanding Users**

Bob Donahue  
WGBH Interactive  
December 2006

### Abstract

There's a lot of data in web logs that tend to go unanalyzed. By tying together referral information and page “hit” data, you can uncover user behaviors that can influence future site development, or prioritize site promotion.

### **1. Introduction**

Web logs are notoriously messy to work with because even a short time span of data can become overwhelming. Yet, they're the basis for determining what's been happening on your site. Typically, when they're analyzed all the focus is on page views but that only gives a one-dimensional view of the site, usually the “hot spots” of which pages are discovered by your users. But that information is clouded by other effects: it might be a page that's heavily linked on the site, but only used by visitors as a “pass through” to some other part of the site that interests them, it might be one whose indexing within search engines is particularly favorable so that it shows up better than other pages on the site, and one would usually expect that the primary nodes of the site (e.g., the main “index.html” page) should receive more hits than the pages at the deepest levels.

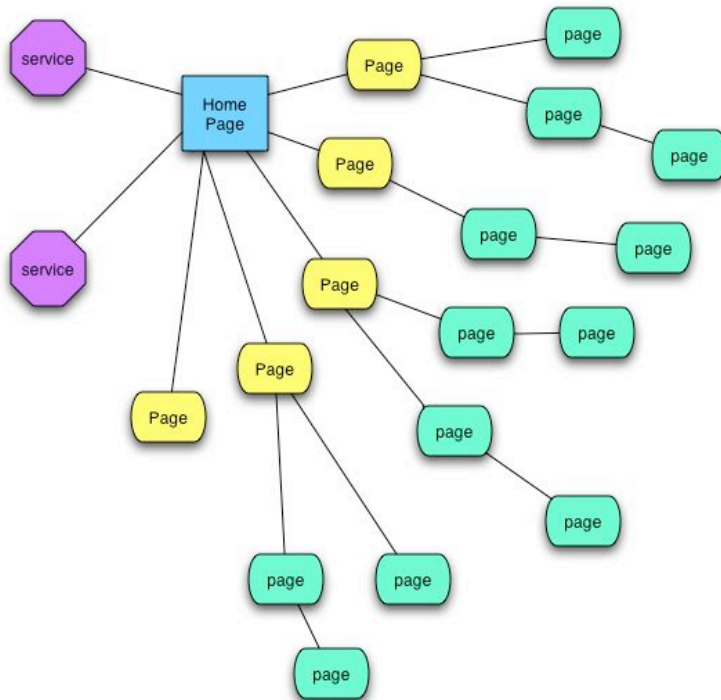
After scratching the surface, other metrics start to become interesting. Entry pages inform you as to what pages are being found from links on external web pages or through search engines. Exit pages are important if there's a series of pages that follow a process such as registration, etc., to learn where “fall out” – when a user leaves the process prior to its completion – occurs.

The next level is intra-site path analysis. Here the range of questions is only limited by the complexity of the site, your goals, and your imagination. Some examples:

- How many users visited page “X” and page “Y”?
- How many pages did it take for users to get to a particular service?
- Do the routes that users take through the site indicate a particular behavior?
  - It is a behavior that you want?
  - If so, is this because of the content?
  - Is it possible that the site layout is causing the behavior?
- Is page design responsible for some user behavior?

### **2. A Look At Site Layout and Paths**

Here's a diagram of a very small and simple web site:

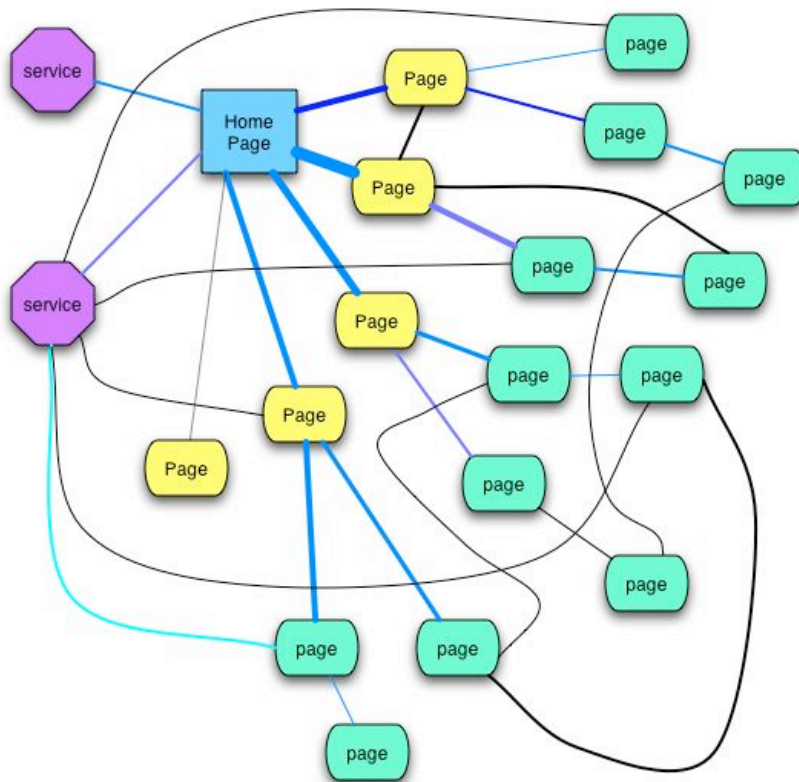


**An idealized web site showing links between pages.**

We have a typical “home page,” two levels of pages below that, and a couple of “services” (e.g., a search service and something else). The lines show the file system hierarchy and let's just assume that for the most part links on pages at least follow this hierarchy although there may be links between pages that “jump” parts of the hierarchy (e.g., all pages might have access to the search service).







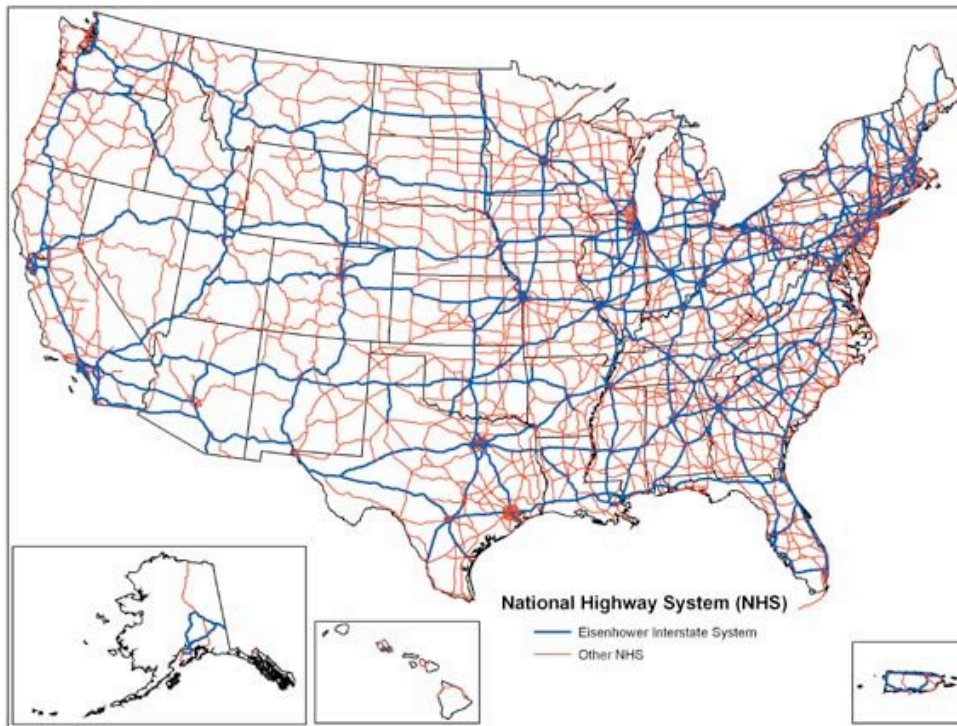
**Many user paths. Thickness ≈ "count"**

Here, I've changed the representation slightly to where the line thickness is a crude indication of the number of times the page from two nodes (in either direction) occurs<sup>2</sup>. As you can see, things quickly become complicated because no matter how you set up your site, you're at the mercy of the users' desires. But this is getting closer to the real situation that paths generate as the effects of bookmarking, offsite linking and cross-site page linking becomes more open-ended. Since many sites have hundreds if not thousands of pages the mapping can quickly become convoluted, more so with the addition of services that can be reached from almost anywhere on the site (such as a typical search engine) to almost anywhere else on the site.

When trying to visualize something of this scale, the following image example comes to mind:

---

<sup>2</sup> Ideally, you'd want to represent this with two lines, one indicating the number of traversals in each direction.



There are definite nexuses of traffic (the metro areas) and certain routes that receive more traffic than others. Yet even if the paths are strung out (in this case across the length of the country in any direction) the actual paths traveled by each user is usually only a small subset of the whole either as a short route or a long trip. In this example there are even disconnected parts of the network (Alaska, Hawaii, and Puerto Rico) that have traffic patterns, and that analogy can also describe some web sites (“internal” sites vs. “public facing sites”).

### 3. The Data Available

Each record in the web server log contains the two things you need to begin looking at user paths: the referrer URL and the destination URL. Typically it’s the latter that is examined for determining path “counts” or a few other metrics. However, the referrer also has quite a bit of information that is frequently overlooked such as entry pages. The count of each referrer/destination pair creates a vector: the direction coming from the pair itself<sup>3</sup> and the vector “length” represented by the number of times it is traversed.

But what is this good for?

#### 3.1. “You Can’t Get There From Here” – Site Navigation Inadequacies

When we developed the site layout for Teachers’ Domain we planned it in a very hierarchical manner with disciplines (e.g., Science) above subjects (e.g., Life Science)

<sup>3</sup> It’s important to remember that in this case going from page A to page B is different than going from page B to page A, just like the separate lanes of travel on a highway.

above topics (e.g., Evolution) and finally subtopics (e.g., Human Evolution). Subtopic pages had a list of resources with a list of multimedia resources. The expectation was that a teacher would drill down to find useful materials. When we entered the pilot program using Omniture, one of the reports I examined was the “Complete Paths” for some of the longer sessions (dozens to a couple hundred of pages viewed).

What I quickly realized was that some user paths were more circuitous than I had anticipated because of the page designs we created under this hierarchy. In particular the following behavior occurred frequently:

- Drill down to a subtopic page of interest
- Look at the first resource on the list
- Hit the back button
- Look at the second resource on the list
- Hit the back button
- Look at the third resource on the list
- ...

So, the number of “counted” page views was really higher than “actual” page views because the subtopic page was reloaded again and again owing to the fact that there was no way to jump to the “next” resource in the list from the resource page being viewed.

In the diagrams above, it would look like a “comb” going down and back between one page and all the pages below it instead of a “fan” starting at one resource page and traversing them directly in sequence. This realization is leading to changes in the page design that will definitely improve the user experience. The “downside” is that the increase in site navigation efficiency will also have the effect of lowering our page views since users won’t have to click so much to cover ground<sup>4</sup>.

### **3.2. Popularity Doesn’t Always Follow Promotion**

One of Omniture’s reports is a ranking of Entry Pages. The canonical “wisdom” when setting up a site is that most visitors will come to you from your home page and surf your site from there. To that end, a lot of effort gets placed on the design of the home page to help users locate what you’ve decided is the most important facets of your site. As usual, users tend to “have it their way” and it might not always follow your initial plans.

Another TD anecdote: our resources were divided among four grade bands based in part on how they were organized across many state curriculum standards: K–2, 3–5, 6–8, and 9–12. [NSES is K-4, 5-8, 9-12] Owing to the availability of particular resource materials

---

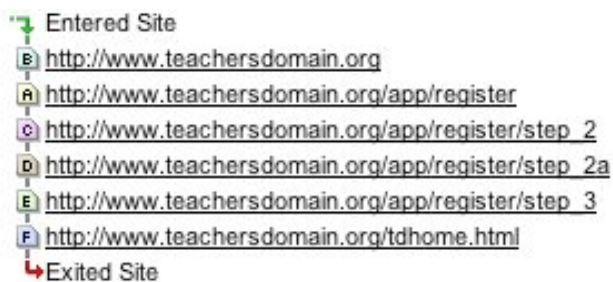
<sup>4</sup> ... once again demonstrating that “raw” counts (here, page views) are problematic: without being extremely certain about the combined effects of many issues (such as the underlying site design), it can be difficult to know what your numbers mean, let alone how those same numbers ought to be compared to numbers from other sites, different sources, etc.

in science and the concentration of interest on certain grade bands – particularly middle and high school – most of the resources ended up in the upper grades. When we looked at the entry pages to TD from search engines, we found something surprising: the requests for K–2 lesson plans in science occurred in much greater frequency than anticipated, showing a surprising demand for materials for those grades and the existence of a previously unidentified Web-active user population (K–2 teachers). This is something that site traffic analysis itself wouldn't have easily provided, since as the number of resources were fewer, there would be less traffic on those sections of the site (fewer “prongs” on the “comb” as described above).

#### 4. Tune In, Log On, Fall Out?

Paths are also important when you've designed parts of the site around the intent that a user ought to completely follow a series of pages. Common examples are registration and surveys that extend over one page.

Again, Omniture's reporting is very helpful about keeping track of this for those parts of your site. In the case of TD registration, here's the sort of path we'd like to see:



although we'd prefer that they don't exit the site immediately after registration. The issue here is that this is only one example of many that “hit” all the pages of the registration process, so what we want is to look at all the user paths, but concentrating only on the registration pages. Omniture's “Fall Out” feature lets us do that. In this case there are three pages that the registration process needs to hit to be successful<sup>5</sup>:

1. /app/register
2. /app/register/step\_2
3. /app/register/step\_3




After step 3, the user either ends up back on the page they were viewing before they registered, or on the TD Home page if they registered from the login page. The resulting conversion diagram looks like this (from 11/1/06 to 12/10/06):

---

<sup>5</sup> The /app/step\_2a page is only necessary if they need to also register their organization.



### Checkpoint Analysis

	Visits		Process	
1.	11,233	100.0%	 <a href="http://www.teachersdomain.org/app/register">http://www.teachersdomain.org/app/register</a>	
			85% Continued	15% Lost
2.	9,535	84.9%	 <a href="http://www.teachersdomain.org/app/register/step_2">http://www.teachersdomain.org/app/register/step_2</a>	
			86% Continued	14% Lost
3.	8,157	72.6%	 <a href="http://www.teachersdomain.org/app/register/step_3">http://www.teachersdomain.org/app/register/step_3</a>	
<b>Total Conversion = 8,157 (72.6%)</b>			<b>Total Fall-out = 3,076 (27.4%)</b>	

and we have an overall conversion rate of ~73%. Since the fallout rate between each step is about the same (~15%) we might improve upon the conversion rate by reengineering the registration process to fewer steps.

## 5. Path Dissection

Omniure also has a much more customizable path analysis tool called “PathFinder.” Here, you have complete control about where to put the checkpoints, and to allow any number of other page visits to occur around them. They come in five styles:

1. What paths users take to get to a particular location;
2. What paths users take after reaching a particular location;
3. What paths users take between two locations;
4. What paths users take on both sides of a location (#’s 1 and 2 combined);
5. Any other combination you dream up (more or less)!

To show an example, let’s go back to the situation described above with K–2 resources.

### 5.1. Preceding Pages and Following Pages

I’ve selected the “Living vs. Nonliving” lesson plan. Running the analysis, we find hundreds of paths (which is to be expected), but ~75% of them are in the top 5 with over half in the #1 path where the lesson plan was the first page visited. This could be from a user’s bookmark, a search result, etc. and verifies what we noticed already. In several other cases, the paths also start with the lesson plan as the entry page, branching off into resource page views, media views, etc. Some paths come after registration (so this was the page the user was on before going through the registration process<sup>6</sup>).

The following pages show similar behavior, with users “checking out” resources linked to on the lesson plan page, registration, media views, etc.

---

<sup>6</sup> ... suggesting that a good follow-up analysis would be to see what types of pages are in this category since they could indicate content or features that are “driving” registration!

## 5.2. “Sandwich” and “Bookend” Reports

“Sandwich” reports have the page of interest in the middle of wild cards (from anywhere, hit the page, go anywhere). These are interesting because they tell you which links on which pages are the most effective at driving traffic to a particular page. Depending on the site layout, navigation, user interface, etc. the results can be quite diffuse if it’s easy to get almost anywhere from anywhere on your site. With the “Living vs. Nonliving” lesson plan, now the top 5 paths only account for ~20% of the paths<sup>7</sup>, but they’re mostly combinations of pages in the same area of the site – the “Characteristics of Living Things” topic where the lesson plan lives. This suggests that some users' sessions remain concentrated within particular parts of the site and that the site design is working adequately (at least for this case).

“Bookend” reports are the opposite: you define two (or more) known pages and allow the user to freely navigate between them, as long as they eventually hit both pages in the order you specify<sup>8</sup>. To look at this, we need to choose another page, so let’s select the search service.

In this case, the top five paths cover ~90% of the cases. Most of them have the user selecting one of the search results (a resource page) that is used as part of the “Living vs. Nonliving” lesson plan, and then going to the lesson plan page. This might argue that the “real estate” taken up on our resource pages listing links to lesson plans using that resource is justified.

**Resource: Animals Making a Living** Recommended for: Grades K-5



Media Type:  
**QuickTime Video**

Length: 2m 15s  
Size: 3.1 MB

[View](#)

[Save to a folder](#)

An animal makes its living by finding food. Plant eaters have a relatively easy time of it, while meat eaters must work a little harder for their next meal. This video segment explores the wide range of food-finding strategies that exist in the animal world and identifies some of the physical and behavioral adaptations that make them effective.

**Background Essay** | [Discussion Questions](#) | [Standards](#)

Different animals have different ways of acquiring the food they need to survive. Animals that consume only plants are called herbivores. This way of life, followed by deer, giraffes, elephants, squirrels, and many other creatures, generally requires relatively little time searching for food (plants are often plentiful) and a lot of time eating (plant tissue contains less energy and is often more difficult to digest than animal tissue, so more

**Topics Covered**

[Organisms and Their Environments](#)  
[Regulation and Behavior](#)

**Lesson Plans Using This Resource**

[Living vs. Nonliving](#)

---

<sup>7</sup> I did make a slight change doing this report in that I also required that the previous pages had to be from within TD (no site entries) since that information is well-established from the previous reports.

<sup>8</sup> This is an important point to remember: going from page A to page B does not necessarily convey the same message as going from page B to page A does.

### **5.3. Wrap-Up**

Of course, this only scratches the surface of the range of usage patterns that can be examined. The scope really is only limited by your imagination and the structure of your own site.