A LIST APART

Beyond Goals: Site Search Analytics from the Bottom Up

by **Lou Rosenfeld** · September 22, 2009

Published in Browsers, Content Strategy, Usability, User Research

Avinash Kaushik demonstrated that site search analytics (SSA) is a powerful tool *(http://www.alistapart.com/articles/internal-site-search-analysis-simple-effective-life-altering/)* you can use to assess customer intent quantitatively. In SSA, as with all flavors of web analytics (WA), you can work from the *top-down*; by starting with clear, measurable metrics based on your organization's goals, you can benchmark and continually optimize the performance of your content and designs. While goal-driven analysis is wonderfully useful, we'll explore a different, "*bottom-up*" approach that relies on pattern analysis and failure analysis to help you understand your users' intent in qualitative ways that complement the top-down approach.

User behavior—it's yours to discover

Rather than measuring performance by key performance indicators (KPI), in bottom-up analysis you "play" with the data to uncover the unexpected: Interesting patterns in the ways people search your site and strange "outliers" that teach you something new about your customers. For example, if you manufacture printers, you might be surprised to learn that your most common search queries are actually for printer drivers, and not product information. Once you realize that existing customers are searching much more frequently than potential customers, you might drastically alter the content you invest in the most.

To understand customer intent, bottom-up analysis is as important as top-down analysis for two reasons:

1. **Analysis always benefits from different perspectives.** No one perspective is complete and authoritative; bottom-up analysis is another lens with which to observe and draw conclusions from your data.
2. **Top-down analysis means measuring only known goals.** Top-down analysis doesn't anticipate the unknowns that arise as your site, your business, your customers, and the world itself change over time. Without bottom-up analysis, you'll miss out on important discoveries that aren't goal-driven.

Additionally, there are occasions where you'll need to rely on bottom-up analytics because top-down analysis won't work. For example:

- **A site may not have clear and obvious goals.** For example, management may not be able to clearly articulate your organization's goals.
- **A site's goals may not be measurable.** It might be difficult to generate useful KPI for your personal website, your daughter's elementary school site, or the local YMCA's site.
- **The act of measurement may not be feasible.** You may not be able to perform measurement because you lack analytics software, time, or expertise.

This is where bottom-up analysis can help: Pattern analysis uncovers trends in the types of information users want. Failure analysis helps you identify the screw-ups that you'd better fix as soon as possible. Let's get started.

Querying your search queries

Bottom-up analysis is simpler than you think: You really do just sift through your data in a variety of ways and wait for interesting patterns and outliers to emerge. For example, examine the top fifty most common search queries. Can you categorize them by topic, or by the types of documents that searchers request? Can you categorize them in some other way?

This really is informal. You don't need to master Excel's most inscrutable formulae, nor do you need a statistics degree. Just dive in and have fun. You can start with your analytics software's basic reports, or by parsing your raw data into a format that you can drop into Excel. As you play with the information, "ask your data" some generic questions:

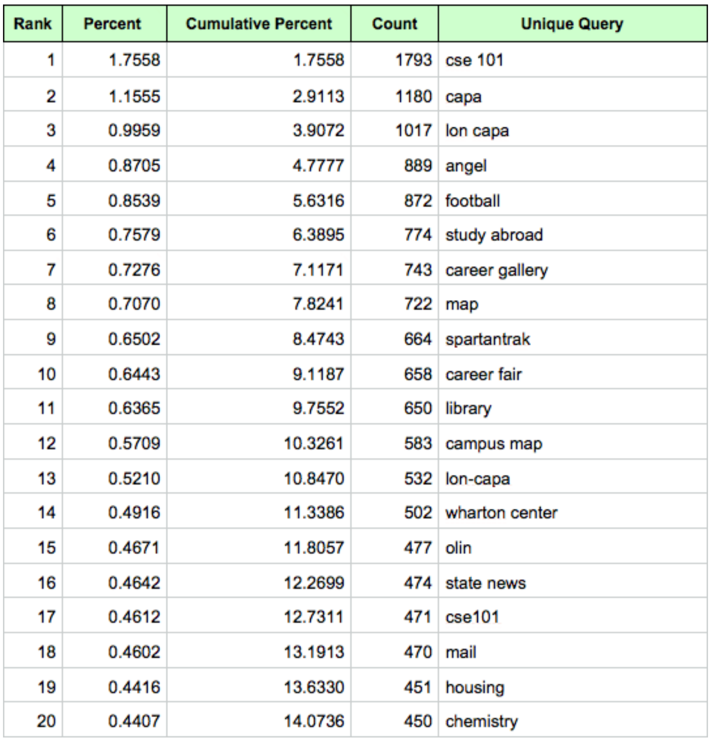
1. What are the most frequent unique queries?
2. Are frequent queries retrieving quality results?
3. What are the click-through rates per frequent query?
4. What are the most frequently clicked results per query?
5. Which frequent queries retrieve zero results?
6. What are the referrer pages for frequent queries?

7. Which queries retrieve popular documents?
8. What interesting patterns emerge in general?

These basic questions are relevant to just about any site, and the answers will often lead you to follow-up questions specific to your site and its users. They're the ideal guide as you confront—and jump into—megabyte after megabyte of search data. And they'll help you with the next steps: pattern analysis and failure analysis.

Pattern analysis

Here's a data sample from the Michigan State University site. Stored in Excel, it includes one week of search queries taken in October and sorted by most to least common:



Rank	Percent	Cumulative Percent	Count	Unique Query
1	1.7558	1.7558	1793	cse 101
2	1.1555	2.9113	1180	capa
3	0.9959	3.9072	1017	lon capa
4	0.8705	4.7777	889	angel
5	0.8539	5.6316	872	football
6	0.7579	6.3895	774	study abroad
7	0.7276	7.1171	743	career gallery
8	0.7070	7.8241	722	map
9	0.6502	8.4743	664	spartantrak
10	0.6443	9.1187	658	career fair
11	0.6365	9.7552	650	library
12	0.5709	10.3261	583	campus map
13	0.5210	10.8470	532	lon-capa
14	0.4916	11.3386	502	wharton center
15	0.4671	11.8057	477	olin
16	0.4642	12.2699	474	state news
17	0.4612	12.7311	471	cse101
18	0.4602	13.1913	470	mail
19	0.4416	13.6330	451	housing
20	0.4407	14.0736	450	chemistry

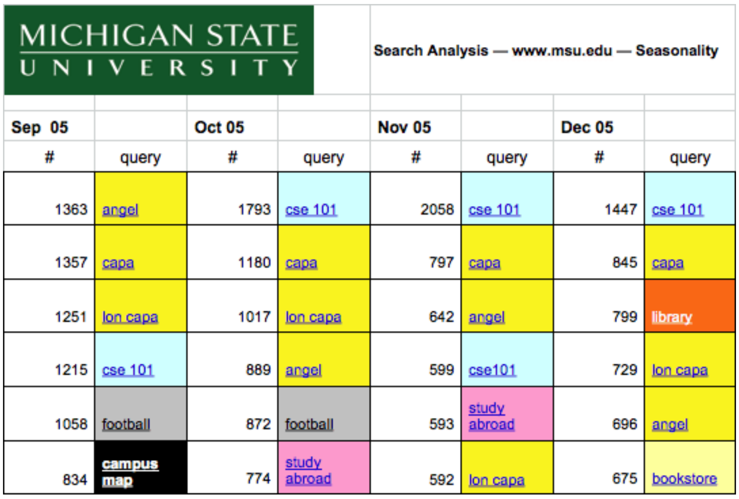
Fig. 1. Michigan State University search results from October, 2006. Figure courtesy of Rich Wiggins.

On just a quick review, some interesting questions emerge:

- Why is the course "CSE 101" the most common query? No other courses crack the top 35. What exactly do users want to know about this course?
- Why are "campus map" and "map" such common queries when the campus map is so clearly displayed on the site's main page?
- Is there a problem with the site's navigation? Or with how the map is displayed? Or maybe there's no problem at all—maybe a lot of users just like to search?
- Why would "housing" rank so highly even though the semester is already underway? Is "housing" queried as frequently at other times of the year? Let's see if the data reveals which documents those who searched for "housing" visited in October when compared with, say, May. What will the differences reveal?

Note that while *none* of these questions have to do with KPI, each is important nonetheless. After all, 2.5% of all searchers sought information about the “lon capa” system. (Lon capa is a course management system.) During this particular week in October, 2.1% of all searchers searched for a variant of CSE 101. Another 1.2% searched for maps. These three queries (and their variants) account for over 5% of that week’s search activity. If you’re the Michigan State webmaster, you should look into how well your search engine supports those searchers, and whether you have content that serves those searchers.

Categorizing these queries will really help you understand the data’s patterns, and you don’t have to be a librarian to do it. All sorts of categorization approaches could be applied; it depends on the patterns that emerge most clearly for you. The following chart shows queries color-coded by category, and mapped out over time. It took about an hour to create:



MICHIGAN STATE UNIVERSITY

Search Analysis — www.msu.edu — Seasonality

Sep 05		Oct 05		Nov 05		Dec 05	
#	query	#	query	#	query	#	query
1363	angel	1793	cse 101	2058	cse 101	1447	cse 101
1357	capa	1180	capa	797	capa	845	capa
1251	lon capa	1017	lon capa	642	angel	799	library
1215	cse 101	889	angel	599	cse101	729	lon capa
1058	football	872	football	593	study abroad	696	angel
834	campus map	774	study abroad	592	lon capa	675	bookstore

Fig. 2. Michigan State University queries color-coded by category, and mapped out over time. Figure courtesy of Rich Wiggins.

Examining query frequency over time introduces another interesting facet: *Seasonality*. Queries that represent systems, (coded yellow), decline over the course of the semester, perhaps as students become more familiar with those systems. (In this context, systems are applications that take you away from the web.) *Maps* (black) are more useful at the start of the semester, the *library* (orange) as finals approach, while *football* (gray) queries decrease as the MSU team spirals downward in another dismal season.

Or, at least, that’s how it seems. Ultimately, analytics tell us *what* is happening, not *why*. After detecting data patterns, we might guess what’s going on with reasonable accuracy. But we can’t know for sure unless we conduct qualitative analysis, such as actual user testing, where we can ask people why they do what they do.

When you’ve got your own search data in front of you, start by asking the following questions. Interesting patterns, trends, and outliers will quickly begin to emerge:

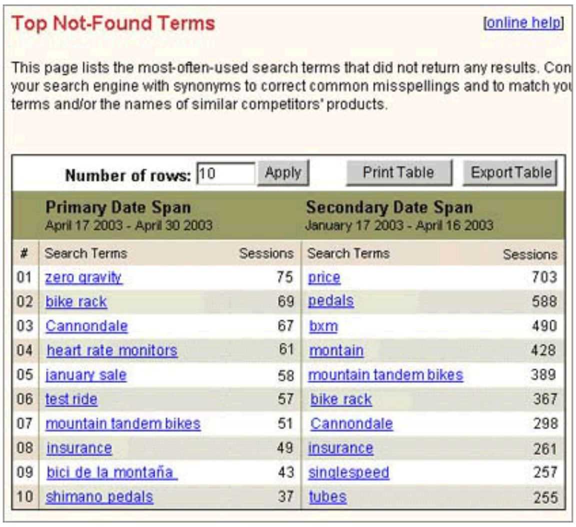
- What are users’ most frequent queries?
- How might we categorize queries (e.g., by task, topic, audience type)?

- What do those categories tell us about our users and what kind of information they need?
- How do timing and seasons affect users' information needs?

Once you start to see some of your site's search patterns, you can use failure analysis to find immediate opportunities to improve the information you provide to your site visitors.

Failure analysis: learn what you need to fix now

Where does search go wrong on your site? If you can analyze your data to find major screw-ups, you'll be able to fix them. To start, simply identify the searches that fail to retrieve any results whatsoever. After all, it's generally a safe assumption that searchers want to retrieve at least one result. Here's an example from a biking products retailer, courtesy of BehaviorTracking.com:



Top Not-Found Terms

This page lists the most-often-used search terms that did not return any results. Con
your search engine with synonyms to correct common misspellings and to match yo
terms and/or the names of similar competitors' products.

Number of rows: 10

	Primary Date Span April 17 2003 - April 30 2003		Secondary Date Span January 17 2003 - April 16 2003	
#	Search Terms	Sessions	Search Terms	Sessions
01	zero gravity	75	price	703
02	bike rack	69	pedals	588
03	Cannondale	67	bxm	490
04	heart rate monitors	61	montain	428
05	january sale	58	mountain tandem bikes	389
06	test ride	57	bike rack	367
07	mountain tandem bikes	51	Cannondale	298
08	insurance	49	insurance	261
09	bici de la montaña	43	singlespeed	257
10	shimano pedals	37	tubes	255

Fig. 3. Top search terms not found.

Here are a few things I observed from playing with this simple report for a half hour:

- "Price" was the top query with zero results between January 17 and April 16. Wow! It's hard to believe that pricing information *wouldn't* be included on the site, but perhaps product prices are buried within each product page. If pricing information is already there, perhaps it's time to redesign the page to make pricing information more prominent.
- Perhaps the retailer, not realizing the potential of the daredevil couples' market, doesn't sell "mountain tandem bikes." If that's the case, it's time to call the manufacturer to place an order. Or, perhaps these bikes are stocked, but the site calls them "tandem trekker bikes." If that's the case, it's time to tweak the product labeling.
- Even though it's not something a bike retailer typically sells, "insurance" is a steadily frequent query. Perhaps there's an opportunity to develop a referral program with a speciality insurer?
- Lots of user misspell "mountain" as "montain." In fact, typos feature regularly in most site-search query logs. Maybe the retailer should turn on their search engine's spell-check feature (or acquire a search engine that supports spell-checking).

Failures take different forms in different contexts. For example, Netflix looks at titles that are in demand: The ones that are most searched and most clicked-through (which they learn about from SSA's sibling, clickstream analysis). Of those, Netflix then examines the titles that are failing: The ones that are least likely to be added to customer queues. They can then examine why—is there enough stock, are there movie genres that they don't carry, or something else?

Failure analysis demonstrates what's going wrong with your site and by extension illustrates SSA's value as a diagnostic tool. For example, let's say you estimate that, based on your data analysis, 8% of your user queries include typos. If you estimate that your users perform searches on half of their visits *(http://www.useit.com/alertbox/9707b.html)*, these numbers reveal a compelling conclusion: Installing a spell-check feature to fix typos could improve the overall user experience of your site by 4%: [8% (searches gone wrong) x 50% (portion of users who search) = 4%].

Four percent might not sound like much, and we can certainly challenge that number. But it may be enough to impress your organization's decision-makers—who may be considering far more expensive and less effective alternatives to installing a spell-checker—such as a redesign *(http://www.slideshare.net/lrosenfeld/redesign-must-die-381947)*. Besides, if you can improve the site four percent here and three percent there, those little numbers start to add up.

Meet in the middle

We've discussed two basic types of bottom-up analytics. The value you derive from each type depends on the data you start with. And that, of course, depends: Your data might be from a text file in a search log, a search engine or analytics tool report, or in some wonderfully flexible database that supports ad-hoc queries and custom reporting. Whatever the case, you should find these Excel-based examples useful—they're low tech, low cost, and most importantly, they expose you to the actual analysis process that your favorite analytics tool may hide.

In fact, be wary of the standard reports that come with your analytics application. They certainly have value, but these reports also provide a false sense of security—as if they were designed with your needs in mind. Nothing could be farther from the truth: Top-down, goal-driven analytics should be centered on *your* KPI, and your organization's goals aren't the same as everyone else's. Similarly, your search query data—and the users, content, and actions they represent—are unique to your context. Top-down and bottom-up analytics will benefit you in different ways, and if you can find a happy middle ground, you'll have an unequaled understanding of your customers' intent.

LEARN MORE ABOUT SITE SEARCH ANALYTICS

Hurol Inan's *Search Analytics: A Guide to Analyzing and Optimizing Website Search Engines (http://www.amazon.com/Search-Analytics-Analyzing-Optimizing-Website/dp/1419626094/ref=sr_1_1?ie=UTF8&s=books&qid=1234543973&sr=8-1)* (BookSurge, 2006) is an excellent SSA resource. My own book, *Search Analytics: Conversations with your Customers* *(http://www.rosenfeldmedia.com/books/searchanalytics/)* (Rosenfeld Media, 2009), co-authored with Marko Hurst *(http://markohurst.com/)*, will be published in the coming months; our book site contains many useful SSA links and resources.

Happily, many of the best experts practicing SSA share their wisdom on their blogs: Avi Rappoport *(http://searchtools.com/)*, Gary Angel *(http://semphonic.blogs.com/)*, Rich Wiggins *(http://wigblog.blogspot.com/)*, and Lee Romero's *(http://blog.leeromero.org/)* sites are all well worth bookmarking. For SSA-related research, check the work of Yahoo!'s Ricardo Baeza-Yates *(http://www.dcc.uchile.cl/~rbaeza/)*, Amanda Spink *(http://www.lboro.ac.uk/departments/ls/people/ASpink.html)*, CMS Watch's Phil Kemelor *(http://www.cmswatch.com/Analyst/19-Kemelor)*, and Jim Jansen *(http://jimjansen.blogspot.com/)* of Pennsylvania State University.

About the Author

Lou Rosenfeld

Lou Rosenfeld is the founder of Rosenfeld Media *(http://rosenfeldmedia.com/)*, co-author of Information Architecture for the World Wide Web *(http://oreilly.com/catalog/9780596527341/)*, and author of Search Analytics for Your Site *(http://rosenfeldmedia.com/books/search-analytics/)*.

MORE FROM THIS AUTHOR

Seeing the Elephant: Defragmenting User Research *(/article/seeing-the-elephant-defragmenting-user-research)*

ISSN 1534-0295 · Copyright © 1998–2017 A List Apart & Our Authors