

Chapter 5

How You Can Use Search Analytics to Improve Your Metadata

Hoarders, Larders, and Metadata

Different species of squirrels take different approaches when they find a tasty nugget they want to hoard. Gray squirrels employ a strategy called “scatter hoarding.” They bury the walnuts and acorns and other nuts they find in many different locations. Red squirrels, by contrast, are “larder hoarders”—they just gather a bunch of nuts and pile them up above the surface on the ground.

It turns out that gray squirrels remember amazingly well where they bury the nuts they find, but they don’t remember perfectly. Those nuts whose location they fail to retrieve often germinate, leading to new trees. It’s a beautiful arrangement: the gray squirrel gets fed during the long winter, and the forest gets a new tree in the spring.

The red squirrels are awful at planting trees. They just leave a pile of nuts on the ground, and those seeds don’t germinate. (Presumably they get a pile of food, unless other critters intervene.)

When humans find a nugget of information they want to save, they too employ different strategies for later retrieval.

In the early days of the Web, most of us used the “bookmarks” or “favorites” feature of our Web browsers to try to capture URLs we thought we might want to visit later. It turns out most of us weren’t very good at it. If we’d been successful gray squirrels, we would’ve carefully put each URL in its own niche. Most of us were red squirrels, and we piled the URLs into an undifferentiated mass.

Then Yahoo came along and tried to organize the Web for us. At first, they were largely successful. We clicked on categories and found useful Web sites. As the Web grew, Yahoo browsing became far less effective. Scale is a scary thing.

If you have a relatively finite number of items to organize—say, several million books for a research library—there are methods you can employ that are time-tested and proven. If you’re a librarian, you can adopt a classification approach such as the Dewey Decimal System or the Library of Congress scheme. You use well-understood ways to describe each book: its author, title, date of publication, ISBN. You, and a community of others (such as the Library of Congress) cooperate to choose a call number that will help visitors to your library find the book right their in proper order on the shelves.

In so doing, you're creating and using *metadata*—data about data. You describe the book from outside the book's own contents. You itemize the characteristics someone might want to use to retrieve the book—our nugget—in the future.

Thoughtful creation and application of metadata can be the difference between a highly effective Web site and a failure. If you operate a Web site that sells a million different auto parts, and your search box doesn't connect your customer who needs that particular hose or bolt with the part they seek, then you've lost sales.

Creation of good metadata can be a very formal process. If you're a researcher seeking the next medical breakthrough, it could be worth millions of dollars to retrieve the right thing at the right time. It might be wise to hire a professional specialist to build a taxonomy—a formal description and organization of the vocabulary of your field. (You might purchase one too¹.) A good taxonomy can help your scientists retrieve that special nugget that helps you design the achieve that breakthrough. This metadata creation could require months of hard work and many thousands of dollars to create—but the payoff could be huge.

But if you're YouTube, del.icio.us, or Flickr, the world is too dynamic. That relatively static taxonomy that you spent \$100,000 to create can't possibly deliver that next new, bizarre video or photograph that millions want to see, right now, today.

So you don't buy it or build it yourself. You let millions of people describe the content they encounter, that they think they might want to retrieve again. They “tag” each URL with the words they'd use if they wanted to retrieve it later. Instead of a taxonomy, they create a “folksonomy”² that will allow millions more find the item—the nugget—they seek. And when your tags you chose personally are amalgamated with those of millions of other folks—guess what, you probably have something much more useful to you than your own bookmarks were.

It's all about metadata. You might spend many thousands of dollars, as libraries and drug companies do, creating specialized high-quality metadata. Or you might let millions of folks do the work for you for free.

¹ There's actually a retailer for taxonomies: Taxonomy Warehouse (taxonomywarehouse.com). Of course, it's inevitable that WalMart will move into this market soon.

² Thomas Vander Wal, coiner of the term *folksonomy*, defines it as such: “Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information.” More at Thomas' site: <http://vanderwal.net/folksonomy.html>

But here's the point: Search analytics can help you understand your customer, your site, and your metadata. When your metadata isn't meeting customer needs, search analytics can show you the way to fix that problem.

Which metadata are *you* likely to be concerned with? Whether you're designing a web site, software application, or some other information system, you're likely grappling with options and content that need to be *described*—options in menus, terms in a taxonomy, links between pages, or tags for individual documents. In this age of information, there's no avoiding metadata. Fortunately, search analytics is an excellent tool to help you develop and tune metadata. That's what we'll cover in this chapter.

Queries and tags: identical twins or distant cousins?

What's the connection between search analytics and metadata? Metadata and search are typically supported by separate technologies, so we often treat them as very different animals. It's always “search *or* browse,” the latter being quite dependent on metadata. But is there really a difference between a search query and a metadata term?

We already know that a query—like “directions”—is an expression of a searcher's information need. Depending on the nature of the Web site, we may have to divine the user's actual intent—does she want directions on how to get to your business, or directions on how to work the photo printer that you sold her?

A metadata term—again, let's take “directions”—is really not that different. It's simply someone else's guess at how a user would express his or her information need or, in the case of folksonomies, it's how the user *did* express his need. They're essentially equivalent.

Well, that's not quite right. Queries and tags aren't always the same. What a searcher might call “directions,” a site designer or information architect might label as “How to Reach Us”. There are many reasons for this discrepancy: the designer has to consider how the metadata term fits in a broader context. “How to Reach Us” might be part of a navigation system made up of consistent labels with intentionally similar syntax, such as “What We Do” and “Who We Work With”. “How to Reach Us” might be part of a taxonomy, with child metadata terms such as “address,” “map,” and—you guessed it—“directions”. There may be internal politics behind the decision to use “How to Reach Us”. And it's fair to argue that our minds work in different ways when we're searching versus browsing.

But the biggest reason these discrepancies arise is that designers don't know what terms users would employ to describe their information need. And that brings us back to search analytics. Because search analytics describes information needs in *users'* heads, it's the best way to eliminate the guessing and connect the dots so that metadata terms better

match with users' terminology. In this chapter, we'll show you how search analytics can help you do a much better job of diagnosing and improving your metadata.

Developing and Maintaining Metadata

Search analytics helps you improve metadata in two ways:

1. Determine *metadata values* (e.g., "W. Shakespeare" or "William Shakespeare") with which to populate those attributes.
2. Determine which *metadata attributes* (e.g., "author" or "playwright") or types are most appropriate to develop.

And, as an added bonus, the same analysis can help you also deduce common content types, like "author biography" or "book review". (We'll cover using search analytics to improve your content in Chapter Six.)

Which Metadata Values?

We'll begin with a sample of queries from a weekly segment of Michigan State University data. In fact, we'll grab the 50 most frequent unique queries from that week; covering over 18,000 individual searches, these top 50 queries represent over 20% of all searches executed that week at MSU, so our sample is at least a reasonably-sized starting point.

It required less than an hour to determine which queries seemed synonymous and tally their respective percentages:

Query	Percentage	Query	Percentage
stuinfo	3.0436	jobs	0.3834
student info	0.4443	employment	0.2659
stu info	0.3047		0.6493
	3.7926	library	0.3601
deans list	0.8509	enrollment	0.3479
dean's list	0.6593	enroll	0.3424
	1.5102	webenroll	0.2449
registrar	0.7446	web enroll	0.2294
mail	0.6171		1.5247
email	0.4532	cemscores	0.3379
	1.8149	degree navig	0.3302
schedule of c	0.564	transfer credi	0.3291
spartantrak	0.503	olin	0.3269
spartan trak	0.2626	angel	0.3191
	1.3296	honors collec	0.3169
bookstore	0.4809	tuition	0.3125
stores	0.2149	transfer	0.2703
bookstores	0.2138	financial aid	0.2692
computer sto	0.3978	academic cal	0.2604
	1.3074	calendar	0.2294
human resou	0.4388		0.4898
HR	0.2327	parking	0.2571
	0.6715	payroll	0.2238
im west	0.4332	dpps	0.2227
study abroad	0.4155	state news	0.2205
campus map	0.3989	grades	0.2116
map	0.3823	nursing	0.2061
	1.6299	psychology	0.205
housing	0.3956	college of ed	0.1994
transcripts	0.3933		
transcript	0.2028		
	0.9917		

Figure 5-1: Queries and query clusters, and their respective percentages of search traffic over a given time period.

Determining preferred terms

Why determine clusters? They help us work through which terms are synonymous, which can be helpful if we need to select a single variant. For example, there is a cluster of terms related to student information (“stuinfo,” “student info,” and “stu info”) that comprises about 3.8% of all the searches that week. Of the three variants, we can tell that “stuinfo” is by far more common than its synonymous siblings, accounting for about 3% of that week’s searches alone. (This isn’t surprising, because at Michigan State University, “stuinfo” happens to be the official name of a branded application.)

Knowing that one variant is far more common than others can be useful if your metadata approach relies on “preferred terms,” single “approved” variants. So we might guess that “stuinfo” is, at least according to users, the best variant to choose to tag documents related to student information. We may also use a preferred term in a relevant document’s title, or in a meta description that’s embedded in the page’s header. If we are developing a site-wide taxonomy, we can typically include only a single, preferred term—and now we know that “stuinfo” is probably the best one.

Working with synonyms

Of course, some technologies and approaches don’t require us to select a single preferred term. For example, if our search systems can index and search metadata, we may want to ensure that all the variants are included. That way, relevant content is retrieved, regardless of which variant is searched.

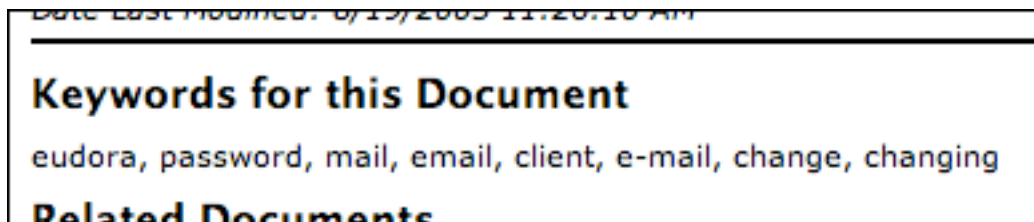


Figure 5-2: “How To Change Your Saved E-Mail Password in Eudora,” a help document that displays synonyms inline to improve its chances of retrieval. (<http://techdev1.acns.msu.edu/drg/article.asp?id=4578>)

The clustering exercise we performed above might help us identify the most common variants to use when populating a search-focused thesaurus or developing search systems that take advantage of *concepts* (large, complex queries that include many variants of common search terms).³

Assessing metadata tone and form

As you read through, cluster, and otherwise play with these terms, you’ll also begin to notice their *tone* or flavor. Do you encounter many abbreviations, jargon, or foreign usage (like *lorry* for truck or, if you’re in the UK, vice versa)? Does the language match what you expected? More importantly, does it match your site’s content? This is a good

³ In the search world, *concept searches* are large, complex saved searches. Search on “child disability,” and you might retrieve documents mentioning that phrase but also “disabled children,” and “disabled youth”. These queries are “souped up” either through editorial effort or automatically; in the former case, humans create concept searches out of popular queries as identified by search analytics.

time to compare common queries with the documents that they'd ideally retrieve. If you're noticing that users tend to abbreviate frequently—say, for example, using the term “im west” rather than “West Intramural Sports Building”—perhaps you should be tagging relevant pages with obvious short forms.

How granular should you go?

This quick dip into the top queries won't help you with a different issue—*term specificity*. How granular should you go? Should your metadata consist solely of general terms like “cats” and “dogs,” or also include narrower ones like “abyssinian” and “yorkipoo”? Terms that are too broad will be of little practical value to users; overly narrow terms also have little value, and will cost you an arm and a leg to develop and apply. And you'll need to work through term specificity if you're developing a thesaurus, a complex set of metadata which includes broader and narrower terms.

To figure out an appropriate level of specificity, you'll need to look at more data, and look at it from different parts of the Zipf Distribution. Mining the long tail will help you get a better sense of more esoteric queries, which often represent different types of information needs than the common queries found in the short head. This is hard work, especially as it's more difficult to find synonymous clusters in the long tail samples. Analyzing individual queries, rather than query clusters, simply takes more time.

A fall back approach is to develop metadata that are a shade more specific than your analysis suggests will be required. This will mitigate some against the likely “short head bias” of your analysis. Even if your terms are initially too specific, users' information needs within a given domain typically evolve to become more specific over time. Erring towards term specificity enables you to stay ahead of the curve for at least a little while.

SIDEBAR: Does your site have an A-Z index? Index terms are simply another form of metadata. Because site indices mix metadata terms of varying degrees of granularity, you might track their relative clickthrough rates. Do you detect that terms of similar degrees of granularity are clicked through most often?

Testing and tuning metadata values

Metadata describe a specific domain, like molecular science or home improvement products. What happens when the scope of that domain changes? The corresponding metadata should, naturally, change to accommodate this new scope, but all too typically lag behind. This is especially true in dynamic, rapidly changing areas, especially those related to technology or the hard sciences. It's simply difficult—and expensive—to keep up. But search analytics offers a variety of approaches to help you evaluate and maintain metadata values; these fall into three categories:

1. Tracking appearances, disappearances, and trends among terms
2. Testing terms by querying them
3. Deriving terms through “reverse lookup”

Tracking metadata trends

Identifying new queries that are trending upward is a great way to keep the scope of your metadata updated. You may be able to identify terms in a particular batch of queries that were never searched for on your site before, and use those terms to develop new metadata values.

Any brand new query—whether searched a thousand times or just once—will register as having infinite growth. Because brand new queries grow at the same rate regardless of their search volume, you’ll want to focus on the more frequently searched new queries. So, if your clothing site has two new queries—*hoodie*, with 71 searches, and *crocs*, with 449—you’ll clearly want to focus your energy on crocs.

But in the case of existing queries that are trending up, it can be a bit tricky to determine which truly merit your attention. Some queries might start small and spike up quickly, while others may already be frequent but are growing at substantial rates. Which are more important?

Term	Last Month	Current Month	Query Growth	% Growth
“2008 marketing strategy”	2 queries	91 queries	89 queries	4,450%
“travel guidelines”	594 queries	712 queries	118 queries	20%

Figure 5-3: Spikes in query activity. Which is more noteworthy?

The more common query certainly deserves attention. It may have a much slower growth rate, but its overall growth (in terms of actual queries) is higher. And it’s already an important query. If there no corresponding metadata value, you might scratch your head and ask yourself why not. If corresponding metadata are already in place, then you don’t have much to do, although something else—unrelated to metadata—might be happening. Seasonal spike perhaps?

The less common query is interesting in a different way. It’s spiked much more quickly, but is it real or an anomaly? Put it on your watchlist, and give it more time to “prove itself” worthy of joining your stodgy, oak-paneled metadata club. Adding a new term to a vocabulary has long term implications, and shouldn’t be done in any manner that hints at flightiness or impulsiveness. So keep your eye on spikes that emanate from the middle torso and determine whether you’ve got a flash in the pan or something of long-term importance. If it’s the former, you might be better of addressing the information need

that it represents with a lighter scale touch, such as implementing a best bet search result to ensure that when it is searched, it retrieves something of value.

Just as a query might suddenly spike, the reverse is true as well: queries can begin to drop out of sight, albeit a bit more gradually than they rise. You'll want to use the same synonym analysis to determine if a more up-to-date variant is available to replace a moribund term.

Keep in mind that eliminating metadata values is a very dicey proposition. You may already have a lot invested in a particular term, in the sense that it's been used to tag a significant amount of your content. Retracting a term will have an impact on each of those documents; without a term you've gotten rid of, will they lose their chances of being found? You'll want to consider replacing stale metadata values with a more viable synonym. Hopefully, you'll have a content management system or some other automated means in place to make such changes globally; otherwise, you'll have a lot of tedious manual labor to struggle with.

Putting metadata through the “query test”

You can also test your metadata by treating them as queries and, well, querying them. Do they retrieve relevant documents? Of course, as mentioned earlier, they're *not* queries, but you'll still find this test instructive—and besides, sometimes they're exactly the same. Here are a few ways you can test your metadata as queries:

1. **Choose a manageable number of common queries**, say the top 25 for a given time period.
2. **See if those queries have synonymous metadata values.** For example, “campus map” may be a common query; its closest metadata equivalent might be “map”.
3. **Note any queries that don't have corresponding terms.** These may indicate gaps in your metadata values that need to be filled.
4. If possible, determine an ideal result for each query. Then **search the metadata values you've identified.** How well do they fare? If a query retrieves zero or few results, the term might merit replacement by a more user-centered synonym.

SIDEBAR: Your site might have important new content. Is it tagged appropriately? Try searching, using the metadata value that you think those new documents should have been tagged with. Are they being retrieved? If not, you might need to do a better job of tagging that new content. And keep in mind that—over time—search analytics will help surface potential new metadata.

Testing metadata values as queries isn't quite as easy as it may sound. You'll need to identify ideal results for each query, and then test performance against that ideal result set. In a small site, that might not be too difficult; in a larger site, especially one that spans many subsites, you'll need to enlist subject matter experts to help you determine

ideal results. As with all other aspects of search analytics, you'll also need to draw the line with your query testing: start with the top 25 queries, then scale up if you have time.

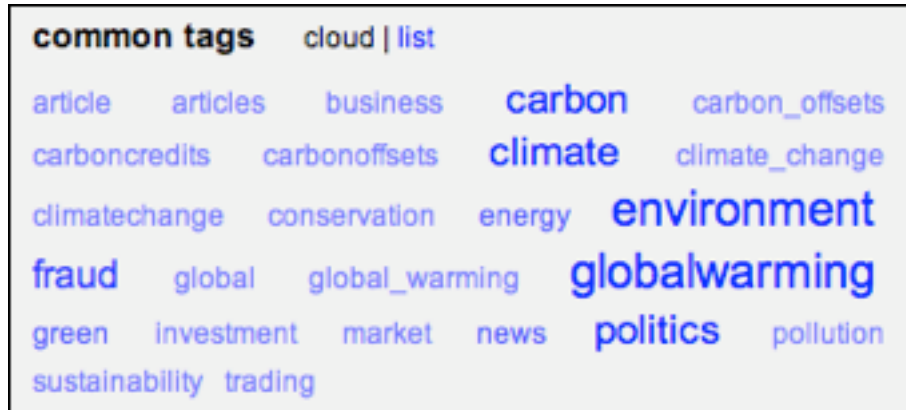
Using “reverse lookup” to identify new and problematic terms

Just as you can test metadata values as queries and see what documents they retrieve, you might consider doing the opposite: start with the documents, and try to determine which metadata values ought to be matched with them. To do this, you'll need an analytics tool that can track the queries that found a particular document.

1. First, **identify “important” documents**. Important can mean different things, such as popular (in terms of showing up in search results or being clicked through from those results), new, or subjectively important according to your business.
2. **List the queries that retrieve those important documents**. You'll have to do some experimenting here, but start with the obvious ones; candidates can come straight from your top queries list.
3. **Determine if there are metadata that correspond with those queries**. If not, then you've identified gaps in your metadata values, or the need to recast existing terms in more user-centered language.
4. If possible, **try this same exercise with new documents**; it will be especially helpful as you try to main the currency of your metadata values.

SIDEBAR: Server log analysis is another way to determine popular documents in terms of pure traffic. You can then check your search logs to see what queries are finding these documents.

ANOTHER SIDEBAR: If you are looking for metadata to describe your public content, you might follow The Financial Times' example. They check to see their pages have been tagged by users of del.icio.us or some other social bookmarking tool, and if so, what terms are used. For example, the article "Industry caught in carbon 'smokescreen'" (April 25, 2007)⁴ was tagged by 84 del.icio.us users with the following metadata:



This is a cheap and easy, if somewhat scattershot, way to harvest candidate metadata values.

As you've gathered, this clustering exercise is by no means scientific, but you can see that even from an hour's effort, you can learn something useful about your metadata values. And of course, you could beef up the exercise's scientific validity by testing more data—by reaching into the middle torso and long tail—and by testing more frequently as well. But spending even an hour every quarter reexamining metadata values through search analytics will benefit your organization—and the people who use its content.

Which Metadata Attributes?

It's challenging to determine which *metadata attributes* to develop (or acquire). There are so many possibilities. For example, take any common object—say, your cell phone, for example, and ask yourself what kinds of metadata could be used to describe it. Here is a handful of the many possibilities:

⁴ <http://www.ft.com/cms/s/0/48e334ce-f355-11db-9845-000b5df10621.html>

Make
Model
Cost
Vendor
Service plan
Warranty length
Color
Style
Size
Weight
Market
Language
Operating system
CPU

Each one of these potential metadata attributes comes with significant overhead in terms of development, application, and maintenance. Choosing the wrong types—or more than you really need—will be an unfortunate, expensive mistake.

Search analytics can help you determine the few best candidates. To illustrate, let's continue our analysis of the MSU data. In the table below, we've listed each query and query cluster on a single line, with its corresponding percentage of total queries.

Then we did something very subjective: for each query or query cluster, we tried to describe the topic of the information that users were seeking. In other words, when someone searches for “transcripts” or “transcript,” we guessed that they might want to complete the *task* of getting a copy of their academic transcript. Or, perhaps, they want to find out which of the university's *services* would help them accomplish this task (in this case, it's the Office of the Registrar). We've noted four potential metadata attributes in the columns to the right:

Query Cluster	Percentage	Potential Metadata Attributes			
		Place	Dept/Prog	Service	Task
stuinfo, student info, stu info	3.79				X
mail, email	1.81			X	X
campus map, map	1.63			X	
enrollment, enroll, webenroll, web enroll	1.52			X	X
deans list, dean's list	1.51				
spartantrak, spartan trak	1.33			X	
bookstore, stores, bookstores, computer store	1.31	X		X	
transcripts, transcript	0.99			X	X
registrar	0.74		X		
human resources, HR	0.67		X		
jobs, employment	0.65			X	X
schedule of courses	0.56				X
academic calendar, calendar	0.49			X	
im west	0.43	X	X		
study abroad	0.42		X	X	
housing	0.40		X	X	
library	0.36	X	X		
cemscores	0.34			X	X
degree navigator	0.33			X	X
transfer credits	0.33				X
olin	0.33	X			
angel	0.32	X			
honors college	0.32		X		
tuition	0.31				X
transfer	0.27			X	X
financial aid	0.27	X	X	X	X
parking	0.26	X	X	X	
payroll	0.22	X	X	X	
dpps	0.22	X	X	X	
state news	0.22		X		
grades	0.21			X	X
nursing	0.21	X	X		
psychology	0.21	X	X		
college of education	0.20	X	X		
Percentage of Queries by Metadata Attribute		4.34	6.35	9.44	9.81

Figure 5-4: Determining metadata attributes from the bottom up, and matching them to queries and query clusters.

So the query cluster of “transcripts/transcript” could be an instance of both a *task* and a *service*. And, by extension, “transcripts/transcript” are possible metadata values for the metadata attributes “task” and “service”. (Note that queries can be categorized under multiple attributes.)

How’d we arrive at these attributes? After starting to review and contextualize the first few query clusters, “service” and “task” rose to the surface right away; these just seemed to be obvious ways to describe these first few clusters. (Remember: we said it was subjective! But sometimes subjective is the best we can do, and it’s better than nothing.) “Place,” “department,” and “program” emerged as we continued our analysis. Once these new attributes surfaced, we had to add new columns and reconsider the first few query clusters and terms to see if these new attributes would describe them as well. Later, we found that “department” and “program” seemed to overlap quite a bit, so we decided to merge them into a single attribute.

SIDEBAR: Want to get users more directly involved in the process? Don't do the categorization yourself; instead, consider a variation on a "closed card sort" exercise⁵, where you present your metadata attributes as categories, and write down the queries and query clusters on index cards. Then ask your test subjects to do the classification for you. You'll always get better data from testing a representative user sample than going by your own opinion.

Once we completed the process of categorizing and classifying each query and query cluster by potential metadata attribute, we totaled the percentages of queries associated with each attribute. For example, to arrive at "place's" 4.34% score, we added the percentages associated with "bookstore, stores, bookstores, computer store," "im west," "library," and other queries that we classified under that attribute.

These scores are especially useful, because they help us to both identify and prioritize metadata attributes. Although there are other factors that we should figure into making decisions on which metadata attributes to support, this exercise provides us with an idea of which attributes to begin with. We have four starting points. "Service" and "task" describe more of the queries, and so seem to promise more value than "place" and "department/program," so if we only have the resources to develop and apply two metadata attributes, those might where we'd begin.

SIDEBAR: Does your analytics tool help you track clickthroughs for each query? You can use that data to further sharpen your prioritization of metadata attributes. Simply replace the percentages of queries with percentages of clickthroughs.

This bottom-up process of extrapolating potential metadata attributes from common metadata values isn't rocket science. Like any other sort of exploratory data analysis, it just requires a combination of iteration and willingness to play with the data a bit. In fact, it only required about an hour to determine these four attributes. And, again, more data—especially including analysis of middle torso and long tail queries—might paint a more complete picture. But even this small amount of effort can really help us leverage real user data to do a better job of selecting and prioritizing the metadata attributes we should support.

Which Semantic Relationships?

The process of divining metadata attributes and metadata values⁶—moving from broad to specific—is really an exercise in understanding the hierarchy that may be implicit in your

⁵ Interested in card sorting? Then read Donna Maurer's forthcoming book on the topic, *Card Sorting*, to be published by Rosenfeld Media in 2008: <http://www.rosenfeldmedia.com/books/cardsorting/>

⁶ And, for that matter, content types, as described in chapter on "SA and Content Development"

queries. It can help you think through the kinds of semantic relationships—broader/narrower terms, related terms, and associated terms—that are the basis of conventional thesauri. Here’s a simple example of the semantic relationships built around a single term, *train*:

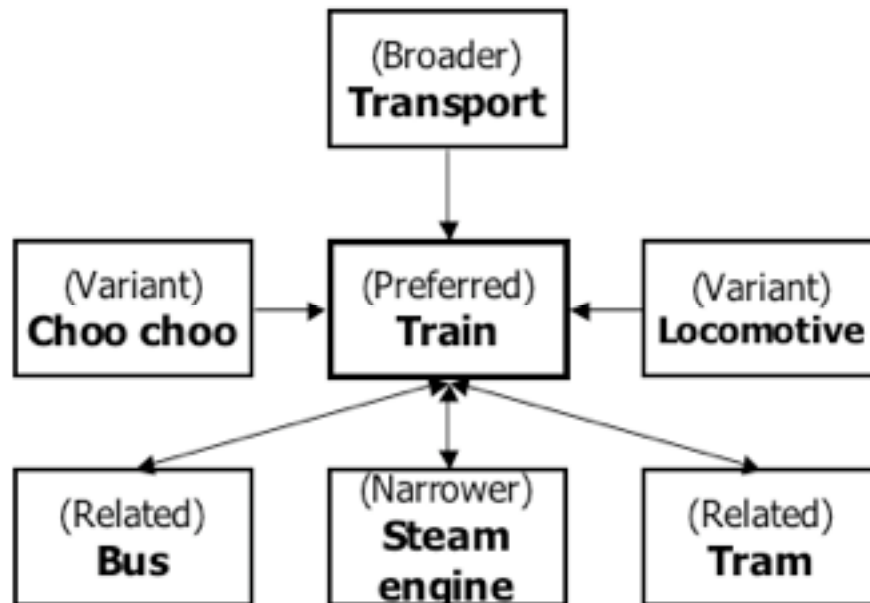


Figure 5-5: Conventional thesauri include a variety of semantic relationships: variant, broader/narrower, and related.

These rich semantic relationships are useful for improving navigation at a page level, helping you easily automatically program links to pages on broader, narrower, similar, or related topics. Search engines can also offer to broaden and narrow results by leveraging thesaural relationships.

For example, if a query retrieves 0 results, we can assume that the searcher would like to broaden her search and net more results. The engine could be taught to automatically insert a broader term or additional variant terms into the query. Leveraging semantic relationships would make it much easier for the searcher to refine and resubmit her query and retrieve more results.

Divination through session analysis

Another interesting (albeit challenging) way to determine broader/narrower relationships is to analyze search sessions. Because human beings are naturally lazy, we often begin with a query that’s broader (and shorter to type) than our actual need. We then enter more specific terms after encountering result sets that are too large and whose content is too general. In other cases, we simply may not know the subject domain well enough to

be sufficiently specific, but as we review search results, we get up to speed and enter more specific follow-up queries. For example:

Searcher...	...and Retrieves
...submits first query: <i>DVD player</i>	291 results
...submits second query: <i>portable DVD</i>	89 results
...submits third query: <i>widescreen LCD portable DVD</i>	9 results

Fig 5-6: Sample of a session where queries become progressively more specific.

After the first query, the user sifts through the first batch of search results and see what kind of language the site uses. She then drops the term “player” in her second query, and gets even more specific for her third and final query. You might extrapolate from this session that *widescreen DVD* and *portable DVD* are subcategories of *DVD player*—especially if you saw this behavior again and again.

If your analytics tool can help you find common sessions paths like this, you might be able to analyze enough session data to get a sense of broad-to-narrow progression. You might also consider sifting through your queries for frequent queries that are quite general. Look at what happens after that initial broad query; do any patterns emerge? Even if the narrower follow-up queries aren’t exactly the same, they might share enough in common—such as syntax—to help you extrapolate a set of sub-categories.

Of course, at the end of any session, we’re usually uncertain that the user was successful (unless the session ends in a successful transaction or other obvious conversion). So while sessions might indicate a broad-to-narrow progression, we don’t really know if that final, narrow term was a good one. If you have time and budget, you might choose to test your conclusions with real, live users; have them list the narrower terms that they would associate with each major query, and see if they match what you’ve come up with through your own session analysis. As always, remember that search analytics tells you *what*, not *why*; and as always, proceed with caution.

The approaches we’ve described so far move from broad to narrow, but you can certainly move from narrow to broad. Categorizing very narrow, specific items—such as queries from the long tail or middle torso—can unearth ever broader categories.

Divination through automation

We’ve already discussed how clustering—even by simple alphabetizing—can make clear synonymous or *variant* terms. Of course, this type of manual effort will only take you so far. If your goal is to truly build a comprehensive thesaurus, you’ll need to use software to look for variant terms across you entire collection of queries. Search engines that support *stemming* will help along the process by identifying terms that share a common

stem or “root,” semantically connecting “engineer,” “engineering,” “engineered,” and “engine” (all share the stem “engin”).

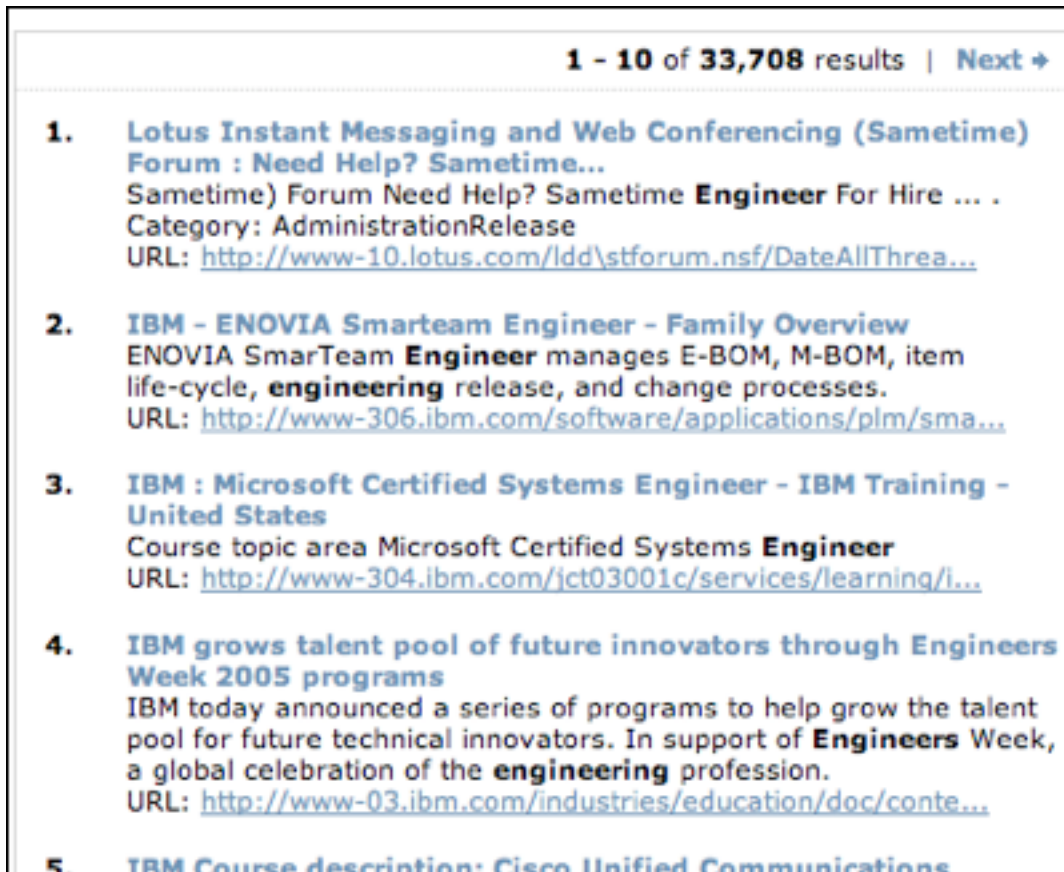


Fig 5-7: A search for “engineering” on the ibm.com site also nets results that include the terms “engineers” and “engineering”.

As with any effort to automate a task that’s generally more appropriate for a human brain, you’ll need to review the results and weed out problems. For example, if your stemmer suggests that “waterski” and “waterboard” as variant terms, you’ll want to decouple them.

If you have more money and more time, you can use software that originates from the artificial intelligence world to determine broader/narrower term relationships and variant terms. These types of relationships are extremely dependent on *context*, something most automated tools aren’t very good at handling. For example, the broader term for “bark” might be “animal sounds” in one context, and “tree anatomy” in another. So don’t forget to budget in significant time to “train” your software to understand your unique context.

Conclusion

There's an interesting paradox here. As information systems become ever larger and more complex, we need quality metadata to navigate our content. Yet it becomes harder to develop, apply, and maintain metadata in the face of all this complexity. Metadata are connecting two moving targets—users and content—and therefore are themselves a moving target. In the face of these challenges, we need better ways to help us diagnose and tune our metadata. Search analytics, an expression of users' needs in their own language, is a great place to turn.

Fred Leise is owed a debt of gratitude for his indispensable help with this chapter.