Coming to terms with tags: folksonomies, tagging systems and human information

Over the past few years we've seen a big movement from hierarchical categories to flat search. Web navigation and email offer prime examples: Yahoo's Directories gave way for Google's search, Outlook's folders are giving way for the search based Gmail. It's far more efficient to come up with and type in a few relevant terms for the page or subject you're looking for than it is to navigate a hierarchy. A hierarchy that may even be built by somebody else, somebody that probably has a different mindset, and therefore categorizes things differently than you would.

Lately, tags have become the hot topic when discussing information organization (or - some might say - the lack thereof). But tagging and flat search are really just two sides of the same coin.

The main reason tagging is so useful is that it resembles and improves search. It assigns relevant terms to web pages, images or email messages, making it easier to find them by typing in one or more of those relevant terms or selecting them from a list. Furthermore, tag based system assign YOUR terms to a resource - based on your mindset - not somebody else's, making them extra useful for managing one's own information. The more terms you assign to the resource initially, the more likely it is that one of them comes to mind when you're looking it up again later on.

But you also want to find new information, not only what you've already found or seen. And again, tagging comes to the rescue. If somebody else has found a term relevant for a page, it's likely that so will you. If that person's mindset (read: use of tags, range of interests and other behavior) resembles yours, the more likely it is to be of use to you. Furthermore, if a lot of people agree on assigning the same tag to a resource, the more relevant we can assume it to be for that resource.

Tags don't tell the whole story

Tags are wonderful, but they won't cure cancer all by themselves.

When tagging a resource - especially in these early days of tagging systems - it is quite likely that you are the first person to tag it. Think as hard as you can and you will still not come up with all the terms that you nevertheless would agree are highly relevant to the subject - terms that you might even think of later when you want to look it up again. It is therefore important to use other methods as well to help assigning relevant terms to the resource, when possible.
Firstly, there are other ways than tagging to help users to identify relevant terms. By allowing them to highlight the snips from the text that they find most interesting, you can give extra weight to that text. Descriptions of the resource, written in a free format are another way. Here you can see an example of a page that has been highlighted using tags, snips and other information from Spurl.net users (another example here).

Secondly you have the author's own information and that should (usually) account for something. In the case of a web page you have the page text and its markup. The markup identifies things that the page author found to be important, among them:

- The page title
- Headings
- Bolded and otherwise emphasized text
- Meta keywords, descriptions and other meta information
- Things near the top are usually more relevant than at the bottom

Analyzing all this text information (both the user's and the author's) in relation to expected word frequency identifies more relevant terms than the tags alone.

And there are even more ways to get relevant terms, link analysis (words that appear in and around links to the document and the relevancy of the linking documents), etc. Basically all the methods that Google-like engines are already using to index pages using spiders and other automated methods.

**Relevant terms are a consensus**

That said, relevant terms for a resource must be seen as a consensus. A consensus between the user himself, other users, the page author and the "rest of the web".
If all of these players are roughly in agreement, all is well and the term relevancy can be trusted. If the information from the page's author is in no agreement with information from the users, it is likely that the author is trying to trick somebody - most likely an innocent search engine bot that wanders by - and the author's information should be given less weight than otherwise.

When searching for something, you don't want to end up empty handed just because nobody had thought of tagging a resource with the terms that you entered. The relevant terms must be seen as layers of information, where your tags and other terms that you have assigned play a central role, then comes information from other users - giving extra weight to those users that are "similar" to you (based on comparing previous behavior), then the page author and finally the "rest of the web".

Trust systems

In addition, you need trust systems to prevent the search from being "gamed". Trust should be assigned to all players in the chain: users, authors and linking web pages. If a long-time, consistent and otherwise well-behaving user assigns a tag to a page, we automatically give it high weight. If a new user includes a lot of irrelevant information on his first post, his trust is ruined and will take a long time to be regained.

Users gradually build trust within the system and if they start to misbehave, it affects everything that they've done before. Same goes for web page authors and linking pages. If your information is in no way in line with the rest of the consensus, your information is simply ignored. Providing relevant information is the only way to improve your ranking. (more in this post on Spurl.net's user forums).

On the author layer, this could even resurrect the long dead keyword and description meta-tags, that are - after all - a good idea, if it hadn't been so blatantly misused.

Coming to terms with tags

The "folksonomy" discussion, so far, has been largely centered on user assigned tags and tagging systems, but these are just a part of a larger thing.

Relevant terms - as explained above - are really what we are after, and it just so happens that tags are the most direct way to assign new relevant terms to a resource. But they are far from the only one. Tagging is not "the new search"; it is one of many ways to improve current search methods. A far bigger impact is coming from the fact that we're starting to see human information - tags being one of many possible sources - brought into the search play.
Hjalmar Gislason is the founder of Spurl.net. The above reflects some of the things his company is doing with the Spurl.net bookmarking system and the Zniff search engine.

Posted by Hjalli at April 13, 2005 12:39 PM