TAG DECAY: A VIEW INTO AGING FOLKSONOMIES

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Terrell Russell
unc@terrellrussell.com

As work continues with folksonomies and databases of tag activity grow large and mature, we will begin to encounter staleness across our tag sets. The problems of agedness and increasing irrelevance can be combated with interesting visualizations and necessary new techniques. From the perspective of each of the other two tenets of folksonomy (objects and people) (Vander Wal, 2007), we can observe how tags are being used over time.

Introduction

Folksonomy, or social tagging, is a relatively new and powerful way to organize information. We've had keywords for some time, but only with the distributed nature of the Internet, more powerful computers, and enough people and bandwidth supplying a rich vocabulary to a variety of resources have we been able to see our collective opinion and observe it in interesting ways.

We can see what many people think of things in near real-time and with countable precision. As tagging technologies spread into more applications and broaden the population of those doing the tagging, we will continue to uncover new ways to slice and dice this extremely interesting data.

But because folksonomy is so new, all the tags and tag datasets we are currently analyzing are still relevant in the time domain. We haven't been doing this long enough to have our tags become stale. But this will surely change.

Background

What happens when we've been tagging for many years - for many decades? Are all the tags that have ever been used to describe an object relevant? Should the searcher be able to define how much an old tag should matter to his investigation? How old is old? And who decides?

This is a struggle that has existed for some time in other fields besides library and information science - to decide how to retire old terms from official usage. Doctors must decide the current terminology for certain diseases; biologists decide how organisms should be classified; they must each work through how terms age and fall out of favor.

What is new here is that this decision-making is getting pushed to the searcher themselves instead of being made by the curator or the panel of credentialed experts. The awareness that terms age out and become less relevant is a burden that adds to the cognitive load of the searcher. Additionally, the searcher must now also divine the reason for this change; to determine whether a term is still relevant, or if it has morphed in some degree into a more current variant.

There are three reasons that a term (tag) can fall out of use with regards to a particular resource: 1) the content (resource) being tagged could have changed, 2) the people doing the tagging could have changed and brought with them a new vocabulary, or 3) the usage of that particular term could have changed over time and no longer means what it used to mean (Coates, 2005). Deciding which one of these has occurred when a term becomes stale is a subtle distinction. We need better tools to help us decide what is actually happening.

Timeline

The recent collective, or aggregate, use of a tag suggests it is still relevant and if its usage drops off over time, we can assume that it no longer has as much relevance to the community doing the tagging. If this trend is being observed for a particular tag, it can be argued that this term is of less importance than it used to be.

Two graphing techniques may give us insight into how this fluctuation over time is occurring. Both involve a timeline with five key points *(Figure 1)*.

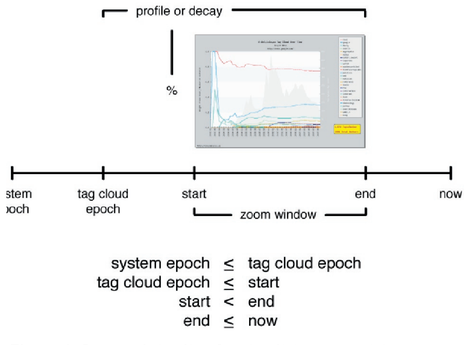


Figure 1. Proposed timeline for visualizing tag activity.

The two end points are fixed and represent when tags were (1) first collected for an object (or used by a tagger) and (5) right now. The second point on the line (2) is where to begin counting the tags and their cumulative use (the zero point). The third (3) and fourth points (4) are the edges of the window of interest (and the time frame we actually graph in Figures 2 and 3). These five points can be displayed with a slider tool to allow for the most interactive and direct feedback when investigating any particular tag dataset. With these five points defined, the researcher can see the tag profile of that object (or tagger) over an arbitrary window of time which allows for a granular inspection of how tag usage has been changing.

The x-axis on the timeline can be represented as both time and as tagging events. The two are directly correlated, but will be variably related depending on the amount of tag activity being investigated. Popular items and prolific people will have more tagging events within a shorter amount of time than those that are not as active. Showing both time and tagging events on the x-axis allows the searcher more context in which to understand what is happening.

Tag Profile

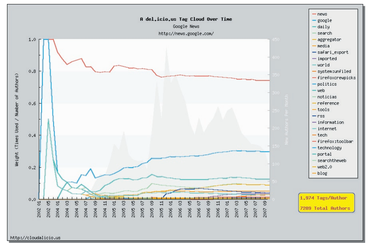
The first technique will allow us to see how an aggregated tag profile for either an item or a person has changed over time. The tag profile is projected from left to right between points (4) and (5) of Figure 1. The y-axis is the percentage of tagging events that used a tag since aggregation began at point (2). Inspecting this graph for diagonal motion, interesting tags can be identified and investigated further (Russell, 2006).

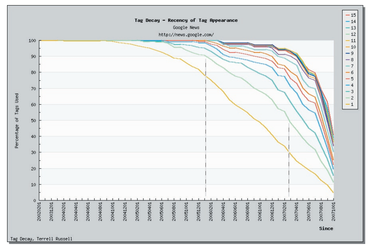
Tag Decay

The second technique allows us to see the aggregate change in a tag profile over time. The y-axis is the percentage of all tags seen since point (2) on the x-axis. And we graph a line each for tags used once, twice, three times, etc. The line formed for common tags (used two or more times) seems to be the most useful, as it is the best general balance between not counting tags used only once (misspellings, special use-cases, personal tagging techniques) and finding the greatest variance before converging at 100%.

Another interesting aspect of this view of a tag cloud's activity is how long ago the common tags reached certain thresholds (50%, 75%, 90%, 95%, 100%). For example, in Figure 2b below, 50% of the "common" tags (second line from the bottom) have been used in the last seven months. 90% of the "common" tags have been used in the last 21 months.

Communities, or populations, with more static vocabularies might go much longer before 50% of its tagging space has decayed through lack of recent use. It also seems fair to assume that the most recently used tags are statistically at the head of the histograph of tags most used. This allows us to suggest that the tags we're no longer seeing recently are in the long tail and might be good candidates for trimming from an official curated vocabulary.





Figures 2a and 2b. Tag Profile and Tag Decay for "Google News". The top three applied tags are "news", "google", and "daily", respectively. 50% of the "common" tags (used two times or more) have been used in the last seven months. 90% of the "common" tags (used two times or more) have been used in the last 21 months.

Future Work

These techniques should allow for interesting comparisons across genres of items as well as different groups of individuals (taggers). Do sports-related web sites have a shorter halflife in terms of the tags being used to describe them than news web sites? Do books from the 18th century have a more stable vocabulary attributed to them than movies produced in the 21st century?

Both techniques could be used for single items and taggers (as above), but could also be used for multiple aggregated items or multiple aggregated taggers at once. This could also be used to determine similarity of a single object or person to a group.

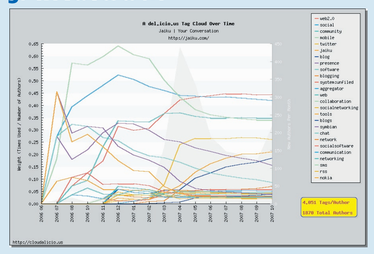
References

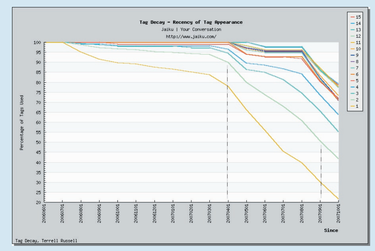
Coates, T. (2005). Two cultures of fauxonomies collide...
http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide.shtml

Russell, T. (2006). Cloudalicious: Folksonomy Over Time.
Proceedings of the 6ACM/IEEE-CS JCDL. (pp. 364-364) New York: ACM.

Vander Wal, T. (2007). Folksonomy :: vanderwal.net. http://vanderwal.net/folksonomy.html

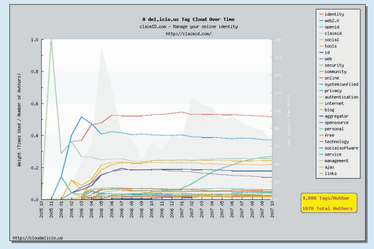
JAIKU.COM

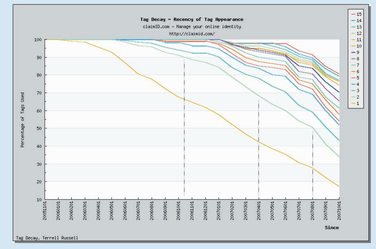




Figures 3a and 3b. Tag Profile and Tag Decay for the web site "Jaiku". There are lots of diagonal lines in this tag profile. This tells us this tag space is very unstable and still in flux. After the large number of tagging events in March 2007, the "presence" tag has been steadily losing ground while "blog" has been gaining. This is most probably due to the different demographic using Jaiku - the people have changed, so the vocabulary has changed. We can see the effect of the March boost as an elbow in the Decay plot as well. The activity has been more fevered since March and we note this as a more rapid churn of older/original terms (a steeper plot). 50% of the "common" tags (used two times or more) have been used in the last month. 90% of the "common" tags (used two times or more) have been used in the last six months.

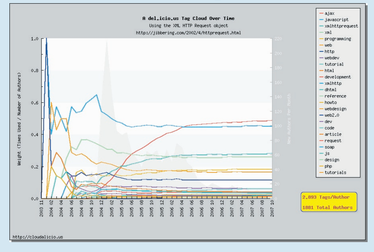
CLAIMID.COM

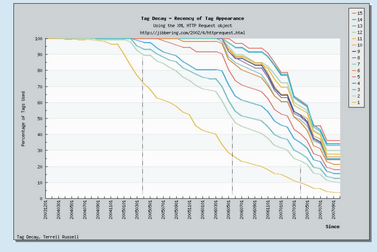




Figures 4a and 4b. Tag Profile and Tag Decay for the web site "claimID.com". The tag profile shows a change in content at the site. In late 2006, claimID began to position its service as an OpenID provider as well as a tool for online identity management. This is apparent with the strong diagonal rising from early 2007. The tag decay shows 50% of the "common" tags (used two times or more) have been used in the last two months. 90% of the "common" tags (used two times or more) have been used in the last twelve months. This suggests that the descriptive nature of the tags being used at del.icio.us is more stable for claimID than for Jaiku. 68% of the "common" tags for claimID have been used in the last six months.

THE RISE OF AJAX





Figures 5a and 5b. Tag Profile and Tag Decay for a web page discussing what would later be termed "Ajax". The term had not been coined when this page was first published - it was a technology without a name as it was a combination of javascript, the browser page, and an xmlhttprequest. The content of the tagged web page had not changed, but the diagonal is clearly rising from the end of 2004. 50% of the "common" tags (used two times or more) have been used in the last 17 months. 90% of the "common" tags (used two times or more) have been used in the last 31 months. Only 24% of the "common" tags have been used in the last six months. The language around this page has settled dramatically compared to Jaiku and claimID.