Harnessing Folksonomies for Search

Eetu Mäkelä

Semantic Computing Research Group (SeCo),
Helsinki University of Technology (TKK), Laboratory of Media Technology
University of Helsinki, Department of Computer Science
eetu.makela@tkk.fi
http://www.seco.tkk.fi/

Abstract. This paper analyses folksonomies, an emergent web 2.0 technology. Folksonomies are found to be primarily a social dynamic phenomenon, and several key tensions are hypothesised that keep the folksonomy community vibrant. Strengths and weaknesses of folksonomies are analyzed w.r.t applicability to browsing and search, and suggestions are given on how to alleviate search problems by bringing in additional semantics into folksonomies, while trying to avoid upsetting the delicate social balances discovered.

1 Introduction to Folksonomies

Folksonomies are a web 2.0 emergent phenomenon, popularized on sites such as del.icio.us[1] and Flickr[2]. A short definition of folksonomies (paraphrasing the definition in Wikipedia[3]) is that they are collections of open-ended tags given by users to content in order to categorize it.

Here, open-ended means that there is no fixed vocabulary nor are there generally any restrictions on what tags a particular object can be given. In virtually all current implementations, this means that tags are unconstrained textual labels typed in by the users.

There is a dynamic tension between private and public in folksonomies. Most folksonomies originate on a need to tag items for one self, but because these tags are shared, also contain or develop an explicit social nature. For example del.icio.us advertised itself in the beginning as an Internet-based bookmark-managing software, and at the time of writing, the tags "toread" and "wishlist" are still among the most popular. In the photo-sharing site Flickr, among the most popular tags are for example "friends", "family" and "me"[1]. Still, even in these private tags there is food for general interest. Which items do people want to read? What kind of people want to find photos of themselves?

Only in the explicitly social aspects of folksonomies are the interesting qualities, the various emergent social behaviours found. Before going deeper into those

[1] http://del.icio.us/
[2] http://www.flickr.com/
[3] http://en.wikipedia.org/wiki/Folksonomy

however, it should be noted that there are multiple kinds of folksonomies with different properties with regard to social dynamics. In a broad folksonomy like del.icio.us, a website is typically tagged by hundreds of users, while in Flickr, a narrow folksonomy, a picture is typically tagged only by its owner[2]. Typically, the social dynamic evolution evident in folksonomies is more prevalent and immediate in broad folksonomies, as feedback mechanisms show what other people have tagged a particular item. However, also in Flickr there is a drive for tag consensus and exhaustive tagging, as users try to make their photos as enticing as possible for peer viewers and commenters.

2 Social Dynamic Properties of Folksonomies

Folksonomies are often contrasted with expert-created taxonomies[3, 4]. While much of the magnanimity against taxonomies is misplaced (arising from the tradition of taxonomies defining a single hierarchy of singular classifications, something no longer pertinent in current ontological thinking), folksonomies do provide some clear benefits when compared to controlled vocabularies. One is ease of tagging. It is much easier for an average user to tag with one's own free-form vocabulary than to get a good enough grip of a controlled, probably hierarchical taxonomy.

Another benefit of folksonomies with regard to taxonomies is that taxonomies usually only provide a single viewpoint to data. In contrast, in a folksonomy the popularity of tags generally exhibits the power law curve common to web 2.0 social phenomena[5, 6]. This means that a folksonomy does create a relatively convergent common vocabulary based on popularity, but also caters to the long tail, individualists and smaller communities with distinct viewpoints to the content, who can still continue to apply whatever tags they like[6, 1].

A second benefit over taxonomies to come from the social nature of folksonomies is the ability to quickly and dynamically adapt to changes in user vocabulary. There is already much hard evidence of this in the del.icio.us tag history graphs available from cloudalicio.us[4]. As an example, in figure 1 is a graph detailing the relative frequency of different tags given to a tutorial article on asynchronous javascript and the XMLHttpRequest -object. From the graph, it can clearly be seen how the term "Ajax" quickly catches on in the community, as a concise term for this core web 2.0 enabler[7]. Another example of dynamic social self-management can be seen by comparing the prevalence of the use of the plural tag "blogs" versus the singular "blog" over time. In figure 2, it can clearly be seen that in the fall of 2004, a dynamic change of policy sweeps over the del.icio.us tagging community, moving to favour the singular form over the plural [8].

The previous examples of tags bring to light an important property of folksonomies: as tagging is kept on purpose totally free of restrictions, there is no consensus on what the semantics of tagging an item by a label mean. Tags can

[4] http://cloudalicio.us/

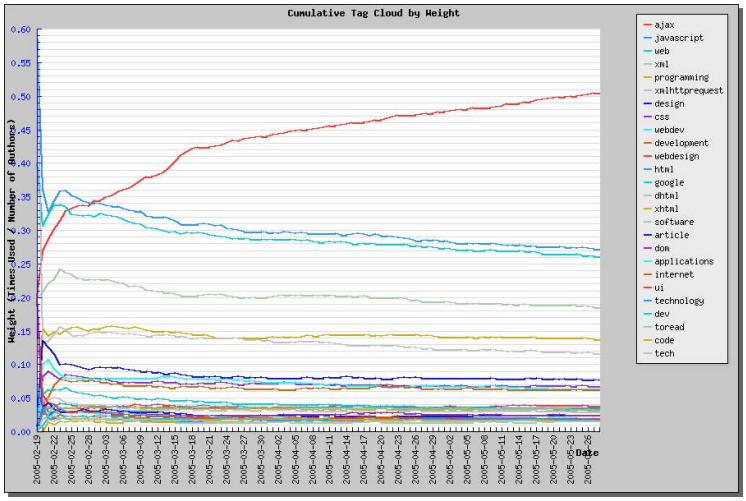

Fig. 1. Emergence of the tag "ajax" in tagging a particular site on del.icio.us

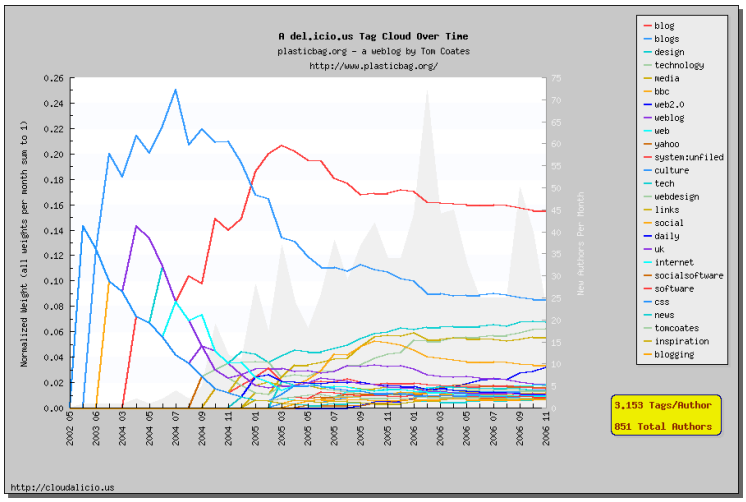

Fig. 2. Emergent movement from use of the plural tag "blogs" to the singular "blog" on one site in del.icio.us

concern properties of an object (this car is "blue"), or properties of the object with regard to the tagger ("wishlist" this car)[5]. In addition, tags are applied indiscriminately to both the form of a presentation (e.g. "Canon EOS 350D", "powerpoint_presentation") as well as the content (e.g. "sunset", "sales_figures"). Taken into a more general level, there simply is no consensus on how to tag. Most users tag "me" instead of "Eetu" and while there are temporal trends towards one or the other, the tags "blog" and "blogs" are still in general equally preva-

lent[1, 8]. Folksonomies also usually subsume multiple viewpoints and cultures, so that similar items can be tagged with a mix of both domain expert and common vocabulary, in multiple languages[5, 9, 6].

A further feature of folksonomies, again arising from the drive for serendipity via multiple viewpoints and as care-free tagging as possible is that while tagging may be the central function of a folksonomy-site, there usually is no, or very limited functionality for discussing the tags or the activity of tagging. Instead of a stifling effect, this actually fosters emergent, unexpected communication and community formation through the very limited medium of tagging itself, especially linked with the resources tagged[1].

From particularly ingenious and serendipidous tags and tagging behaviour arise memes, around which loose communities form, often never formalizing themselves. As examples, in Flickr such communities have arisen around the tag "sometaithurts", where a central meme is that people take pictures of pictures in Flickr in a never ending recursion, but the meme is not really fixed, as there is also other content bordering and molding the meme. Another example of loose community formation in Flickr is when a certain user started combining pictures with small, insular pieces of fiction, and started tagging them as "flicktion". In time, other people started doing this too, commenting on previous flicktion with their own. Taken further, in a sense all tags in folksonomies are venues for communication *about* that tag. For example, the images in Flickr tagged "iraq" tell a story of what the Flickr participants want to say about Iraq, what aspect they emphasize.

3 The Strengths and Weaknesses of Folksonomies as Navigational Tools

A common trend in folksonomy-based user interfaces is to provide as many, and especially different kinds of choices for browsing the different axes of the content space. In both Flickr and del.icio.us for example, one can navigate from an object given a particular tag to that tag, and from there on to other related tags, other objects tagged with that tag or users using that tag. And from a user page, one can again navigate to objects tagged by that user, or to tags used by that user, and so on. This functionality, termed pivoting[10], is especially useful for serendipidously wandering the dataset, discovering new things[1]. Only adding to this is the use of variant vocabulary in folksonomies, as people can freely move between and discover new viewpoints to the dataset by moving between related tags or hopping on a new axis by looking at other items tagged with an interesting tag discovered on an object.

While interfaces based on folksonomies are clearly very suited for tasks falling into the browsing and orienteering[11, 12] categories of search behaviour, it is generally agreed[4, 13–15] that severe challenges are faced when creating interfaces based on folksonomies for spot-search type tasks.

The main problem here, again deriving from the core point of tagging freedom is that there are just too many tags without any hierarchical structure, conflicting

basic levels of tag use (tagging a picture "beagle" vs. tagging "dog"[5]), too many viewpoints mangled together, no handling of synonymy or homonymy, nor even handling of misspellings or consensus on using plural or singular form[1, 6]. The capability of folksonomies to handle changes in vocabulary presented earlier only adds to this: even in the scope of a single user, if they change vocabulary or discover a new subdivision of meaning in the tags they use, any earlier items tagged with other semantics are lost to effective search or even result in erroneous results[5].

In the context of hindering search, all the above are collected under the term *meta noise*, meaning merely that the tag dataset is too heterogeneous and error-prone for basing efficient search upon, lowering both search recall and precision.

There is also evidence in folksonomies themselves that this is a real problem for users. In del.icio.us, users have begun establishing structured tagging conventions that resemble hierarchies (e.g. "Programming/C++", "Programming/Java", "Programming/XHTML"[6]) even when little benefit is to be gained from such, as the system architecture doesn't recognize tag hierarchies. In a longer running, more focused community such as fan-fiction authors, it has been noted that vocabulary seems to converge to a more limited set over time[15].

4 Harnessing Folksonomies for Search

Based on the analysis in the previous chapter, and keeping in mind the delicate social balances that keep folksonomies popular discussed in chapter 2, there seems to be a real urgent need for solutions for reducing meta noise, but which at the same time keep the freedom and ease of use so precious to the folksonomy community. In the rest of this paper, some potential approaches are given on how to do just that.

4.1 Tag disambiguation

A major contributor to the meta noise problem are the homonymy and synonymy problems inherent in the simple text labels used as tags. On the semantic web[16], focused on machine-understandable semantics, the issue has been solved by moving from words to concepts, and giving concepts unique identifiers (URIs). However, such unique identifiers are not easy to create, remember, or to type in directly.

If one were then to use URIs for tagging, one would need a highly efficient interface for suggesting existing tags based on (possibly synonymous) labels, creating new tags based on labels as well as an efficient general mechanism for disambiguating between two concepts with the same label.

As a suggestion for an efficient interface for tagging, a semantic autocompletion[17] -based interface should be considered, such as employed in tagging content in the semantic question-answer system OPAS[18], where existing terms are suggested from an ontology, but also new tags can be created just by typing them. As the problem of homonymous labels occurs also in the pure semantic

web context of semantic annotation, approaches to solve this are already being studied. A natural way on the semantic web to disambiguate individuals is to list their differing ontological environment or properties, such as disambiguating people by their middle names or where they. Of course, this requires thinking about the tags not merely as labels, but as entities in and of themselves, with some possibilities to give these identifying properties to the tags. In essence, this approach would turn the tags into a taggable content type all of their own.

For an average user, this need not complicate the act of tagging, if made optional. This way, one could rely on a core group of power users to create needed disambiguating information, but it would then be immediately useful for everyone. A casual user then would still continue to tag with an "ambiguous" identifier, when no disambiguations are available.

However, to make the disambiguation functionality and use of URIs fully efficient, there would need to be some mechanism for evaluating also prior annotations when new disambiguations are introduced, and the handling of how changes in vocabulary are mapped to the concept space.

To an extent, both could be helped through statistical analysis of the tag base and tagging behavior history. There is already much latent information on the semantics of the tags in how they are used[19]. For example, there probably is a much wider dispersion and less convergence in the use of tags like "wishlist", "me" and "toread", as opposed to tagging a site "ajax" or "tutorial". Therefore, one could conclude that the prior refer to different semantic individuals, while the latter are part of a shared universal vocabulary. Also, particularly polysemous words could perhaps be disambiguated in a partly automatic manner by creating a semantic neighborhood for them of other tags used along with them on different kinds of content. In the same way, or using techniques from the information extraction community[], possible synonyms and vocabulary changes could be at least semiautomatically detected, presented to an editor for final approval.

Of course, these statistical algorithms could also be used not to bootstrap and manage a crisp, disambiguated term space, but in and of themselves to provide fuzzy disambiguation functionality in the search context. The potential problem with such an approach is that it is less forthcoming to semiautomatic solutions and human editorial control, and thus the algorithms must themselves satisfy much stronger precision requirements.

4.2 Semantic Relationships

A big contributor to meta noise is the fact that there are no clear semantics on what it means to attach a tag to a content object[5]. A way to counter this would be to move from simple tags to properties. Instead of tagging "Canon EOS 350D", "sunset", one would tag "takenWith: Canon EOS 350D", and "contentIsAbout: sunset". These properties could also be managed collectively as a folksnomy, and with a fall-back default general property for maintaining interoperability with current functionality. It could be hoped, that especially with this fall-back functionality, this procedure of collecting added metadata would not engender too much cognitive friction on the part of the folksonomy users.

Tagging content this way would be an effective way to partition the tag space, which in turn could be used to drasticly diminish the problems of meta noise. It would aid tag disambiguation, and the partitions could be used as different views into the data. These views could then be used to provide pivoting functionality, or be used for example in view-based search[20, 21], a user-interface paradigm with many useful properties for search and browsing in semantically annotated collections.

4.3 Tying Tags to Controlled Search Vocabularies

A simple way to boost search effectiveness in folksonomy systems, also widely posited in literature[22, 15, 13, 19], is to create hybrid systems combining search taxonomies with flat tag data. In practce, there are many ways to go about this.

A simple idea is to take a controlled vocabulary or vocabularies (for dealing with differing viewpoints in data), and linking the individual tags into these vocabularies for use in search. This can also of course be done in a distributed manner, for example by tagging the tags themselves with controlled vocabulary terms. When doing a search and particularly if seeing the search vocabulary (as in for example view-based search), users should be quite able to map their own search need to the common vocabulary. However, in tagging, their primary process of sense-making, they can still use their own vocabulary. On the semantic web, good results have already been attained using this approach in MuseumFinland[23], where different museums used vocabularies of their own, but which were mapped to a common search ontology for visualizing in a virtual exhibition portal.

For creating and maintaining these controlled vocabularies, one could again make use of the implicit knowledge already inherent in the tag database [19]. For this task, there is a whole slew of algorithms[24–26] for learning ontologies and taxonomies from text data, based on various statistical and clustering algorithms. As a singular example, a hierarchy of tags could probably quite reliably be created by looking for instances where a tag occurs very reliably with another, but also without it (e.g. tag "dog" occurs almost always in documents also tagged "beagle", but also elsewhere. Therefore, a dog is probably a broader term for beagle)[27].

5 Conclusions

Folksonomies seem to be primarily a social dynamic phenomenon, creating vibrant, emergent communities based on a variety of tensions. While these social dynamic properties and variables are many, they all seem to stem from a two key properties of folksonomies: first, that they are very easy to use, and second, that they embrace freedom. Factors leading up to these properties are for example the dual private and social nature of folksonomies, as well as limiting communication about the act of tagging, as well as keeping the tagging vocabularies open.

While these properties make folksonomies very usable for browsing, orienteering and pivoting around data, they are all meta noise for search. Based on the suggestions given in this paper, however, it would seem possible to ease search without endangering the properties that make folksonomies function, by selectively adding semantics to various parts of the folksonomy infrastructure. Testing these hypotheses remain a part of further work.

References

1. Mathes, A.: Folksonomies - cooperative classification and communication through shared metadata. In: Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. (2004) http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html [23.11.2006].
2. Vander Wal, T.: Explaining and showing broad and narrow folksonomies. http://www.personalinfocloud.com/2005/02/explaining_and_.html (2005) [23.11.2006].
3. Shirky, C.: Ontology is overrated: Categories, links and tags. http://www.shirky.com/writings/ontology_overrated.html (2005) [23.11.2006].
4. Van Damme, C.: Folksonomies and enterprise folksonomies. Master's thesis, Vrije Universiteit Brussel (2006) http://www.statbel.fgov.be/studies/ac605_en.pdf [23.11.2006].
5. Golber, S., Huberman, B.A.: The structure of collaborative tagging systems. Technical report, Information Dynamics Lab, HP Labs (2005) http://arxiv.org/pdf/cs.DL/0508082 [23.11.2006].
6. Guy, M., Tonkin, E.: Folksonomies: Tidying up tags? D-Lib Magazine **12**(1) (2006) http://www.dlib.org/dlib/january06/guy/01guy.html [23.11.2006].
7. Speroni, P.: Tagclouds and cultural changes. http://blog.pietrosperoni.it/2005/05/28/tagclouds-and-cultural-changes/ (2005) [23.11.2006].
8. Coates, T.: Two cultures of fauxonomies collide... http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide/ (2005) [23.11.2006].
9. Van Dicjk, P.: Emergent i18n effects in folksonomies. http://poorbuthappy.com/ease/archives/2005/01/15/2419/ (2005) [23.11.2006].
10. Vander Wal, T.: Folksonomy to improve IA. Presentation at Oz-IA Conference, Sydney Australia (2006) http://s3.amazonaws.com/2006presentations/OZIA/Folksonomy_for_IA.pdf [24.11.2006].
11. Sellen, A., Murphy, R., Shaw, K.L.: How Knowledge Workers Use the Web. In: Proceedings of the SIGCHI conference on Human factors in computing systems, CHI Letters 4(1), ACM (2002)
12. Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R.: The perfect search engine is not enough: a study of orienteering behavior in directed search. In: Proceedings of the Conference on Human Factors in Computing Systems, CHI. (2004) 415–422
13. Wu, H., Zubair, M., Maly, K.: Harvesting social knowledge from folksonomies. In: HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, New York, NY, USA, ACM Press (2006) 111–114

14. Rosenfeld, L.: Folksonomies? how about metadata ecologies? http://louisrosenfeld.com/home/bloug_archive/000330.html (2005) [23.11.2006].
15. Lawrence, K.F., Schraefel, M.: Freedom and restraint: Tags, vocabularies and ontologies. In: Proceedings of the 2nd IEEE International Conference on Information & Communication Technologies: From Theory to Applications. (2006)
16. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (2001) 34–43
17. Hyvönen, E., Mäkelä, E.: Semantic autocompletion. Unpublished (2005)
18. Vehviläinen, A., Hyvönen, E., Alm, O.: A semi-automatic semantic annotation and authoring tool for a library help desk service. In: Proceedings of the first Semantic Authoring and Annotation Workshop, ISWC2006. (2006)
19. Niwa, S., Doi, T., Honiden, S.: Web page recommender system based on folksonomy mining for itng '06 submissions. In: ITNG, IEEE Computer Society (2006) 388–393
20. Mäkelä, E., Hyvönen, E., Sidoroff, T.: View-based user interfaces for information retrieval on the semantic web. In: Proceedings of the ISWC-2005 Workshop End User Semantic Web Interaction. (2005)
21. Mäkelä, E.: View-based search interfaces for the semantic web. Master's thesis, University of Helsinki (2006)
22. Lambe, P.: How to kill a knowledge environment with a taxonomy. http://www.greenchameleon.com/gc/blog_detail/how_to_kill_a_knowledge_environment_with_a_taxonomy/ (2006) [24.11.2006].
23. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: Museumfinland – finnish museums on the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web **3**(2–3) (2005) 224–241
24. Buitelaar, P., Cimiano, P., Grobelnik, M., Sintek, M.: Ontology learning from text, slides from a tutorial at ecml/pkdd (2005) http://www.aifb.uni-karlsruhe.de/WBS/pci/OL_Tutorial_ECML_PKDD_05/.
25. Buitelaar, P., Cimiano, P., Magnini, B. In: Ontology Learning from Text: An Overview. IOS Press, Amsterdam, The Netherlands (2005)
26. Cimiano, P.: Ontology Learning and Population: Algorithms, Evaluation and Applications. PhD thesis, University of Karlsruhe (2005 (forthcoming))
27. Cimiano, P., Staab, S., Tane, J.: Automatic acquisition of taxonomies from text: Fca meets nlp. In: Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia. (2003) 10–17