December 2006

The Semantics of Folksonomies: The Meaning in Social Tagging

Csaba Veres
Norwegian University of Science and Technology

The Semantics of Folksonomies: The Meaning in Social Tagging

Csaba Veres
Department of Computer Science
Norwegian University of Science and Technology
Csaba.Veres@idi.ntnu.no

ABSTRACT

In this document we review some key observations about the use of naive user tagging in organizing electronic resources, and the possible advantages they display over rigid classification. We report on aggregated views of communal tags and review arguments concerning the relationship between folksonomy (the emergent process of communal classification) and formal ontology, rejecting the strong view that faces them against each other as opposing solutions. Finally we present a novel framework for abstracting latent structures in folksonomic categorization in a way that facilitates their conversion to formal ontologies. This should enhance their usefulness in integrating resources from different sources.

Keywords

Folksonomy,ontology, semantics, social intelligence, social tagging

INTRODUCTION

The rapid growth in popularity of web sites employing "user tags" has been phenomenal, with the appearance of popular sites like del.icio.us (http://del.icio.us/), Flickr (http://www.flickr.com/), and Technorati (http://www.technorati.com/). Yahoo has jumped right in with the development of My Web 2.0 Beta (http://myweb2.search.yahoo.com/myresults/faq). In addition, companies like IBM are investigating the use of bookmarking services within their intranet. The basic idea of tagging is really not different from the using completely non restricted keywords to label documents. An important advance, however, is the tagging of specific resources by multiple users which results in each resource acquiring a *tag cloud* which is a description of the evolving set of terms being used to describe a resource, together with the popularity of those terms. Tag clouds are often displayed in lists with different sized fonts representing their popularity, as illustrated in figure 1. This mass action of collective tagging has an important side effect in the emergence of folksonomies, or naive systems of classification that congeal from the mass actions of the users. Folksonomies are a collective classification scheme for resources, which can challenge the role of established taxonomies for organizing resources.

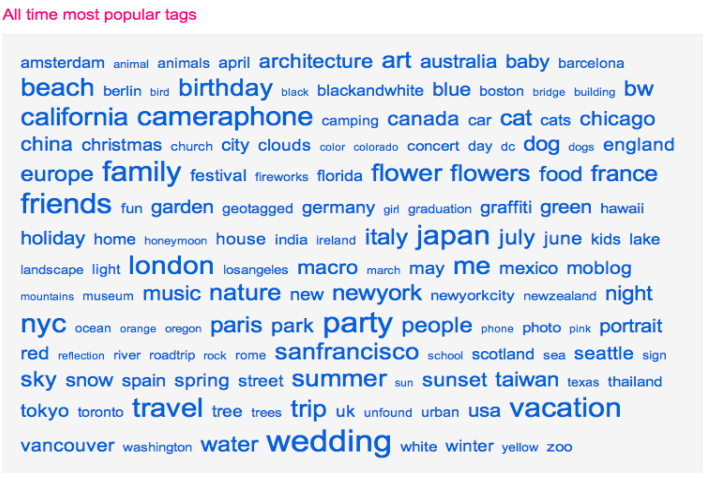

Figure 1. Tag Cloud from Flickr

The idea of cheap, low entry cost user tagging clearly has much to promise, and has been vaunted as a superior way to manage and store information content. A single document can be "filed" in several places because tags can refer to different facets of organization, so users don't need to decide which is the "right" classification. Another powerful consequence of this storage and retrieval paradigm is being heavily exploited in future visions of services like Yahoo's My Web 2.0. The idea is that content of all sorts can be united through the use of tags. For example in writing for this conference I used the on line Word processor at www.writely.com, and tagged the document {paper, AMCIS2006}. But I also tagged the conference web site on delicious with {conference, AMCIS2006}. After I return from the conference I will put photographs on my Flickr web site and label them {Mexico, Acapulco, conference, AMCIS2006, pictures}. It is obvious that if some web service can have seamless access to these different web services, I can easily collect all information relevant to AMCIS2006 if I wish. Or, I can filter them so that I only have my own personal items concerning the conference, and not the official web site, for example. Clearly this is a technology worth thinking about.

But whatever the practical usefulness of user tagging turns out to be, one issue of theoretical interest concerns the nature of the meaning constituted by the tags. That is, are tags fundamentally different from formal categories? Are taggers performing an activity that is fundamentally different from the sorts of activities that are performed in the act of formal classification? This is an important question because it impacts on the way that the tags themselves can be reused in a range of applications. For example we already saw a way that tags could be used in novel ways to link together content residing in different applications, but could they also be used in applications that relied on more traditional, formal category structures? This question has been the subject of much heated debate in the popular literature, mainly on web blogs maintained by IT professionals of one sort or another. Clay Shirky posted a much cited blog in which he criticizes formal classification schemas, lumped by him under the rubric of *ontologies*, and expounds the virtues of free form tagging systems such as *folksonomies*. According to this extreme view the two are entirely different sorts of thing with no overlap in semantics. Ontologies are formally defined logical axioms that can be used to describe content, whereas folksonomies are a "collaborative but unsophisticated way in which information is being categorized on the web" Wikipedia (http://en.wikipedia.org/wiki/Folksonomy). Ontologies attempt to define a rigid, pre-conceived view of the universe on classification schemes whereas folksonomies avoid such errors because they are collaboratively developed in a democratic effort.

But we find arguments about the differences between ontologies and taxonomies to be hopelessly confused. For example within the same source, the Wikipedia entry claims that collaborative folksonomies have the outcome that "*a metadata vocabulary can then be created by democratic effort and the data strength and relevance is improved by collaboration, without the need for a controlled vocabulary to be defined initially. This avoids the inherent errors and potential inaccuracies in a single user-generated folksonomy.*" The main advantage appears to be the highly collaborative nature of the folksonomy development, which helps with quality control. But then there is this, a few lines later: "*In contrast to top-down, authoritative systems of formal taxonomy, folksonomic categories may strike those of a formal turn of mind as hopelessly idiosyncratic, but therein lies their value: a folksonomic category arises from an individual's engagement with the tagged content, such that the created category is simultaneously personal, social, and (to some degree) systematic, in an imperfect and provisional way. Folksonomies therefore convey information on multiple levels, including information about the people who create them, and they therefore invite human engagement. If you agree with somebody's classification scheme, no matter how bizarre it might seem to others, you are subtly but strongly encouraged to explore other objects that this user has tagged.*" But this is an entirely different matter because now the notion of a "*metadata vocabulary created by a democratic effort*" is completely abandoned and replaced by a system that is only "*to some degree systematic*", "*imperfect and provisional*", and even "*bizarre*". Perhaps this stark contrast reflects definitions contributed by different authors at Wikipedia. However, there are also signs of the confusion in the writings of Tim Vanderwal, the originator of the term "folksonomy": "*The folksonomy is a means for people to tag objects (web pages, photos, videos, podcasts, etc., essentially anything that is internet addressable) using their own vocabulary so that it is easy for them to refind that information again. The folksonomy is most often also social so that others that use the same vocabulary will be able to find the object as well. It is important to note that folksonomies work best when the tags used to describe objects are in the common vocabulary and not what a person perceives others will call it ...*". So the vocabulary is simultaneously (i) personal and owned by an individual, (ii) social and (iii) not perceived by the owner with the intent that the vocabulary will be used by others. But these three statements appear contrary: they cannot all be simultaneously true, assuming that (non-autistic) humans possess a theory of mind which allows us to ascribe mental states to fellow humans (Baron-Cohen, 1997).

In this paper we will attempt to clarify some of the confusion by considering the arguments against the available evidence, and introducing a novel analysis which goes a long way toward an understanding of tagging behavior. But we note that the structure of the paper is targeted towards the special interests of the track, the emergence of social intelligence. We will therefore focus on the evidence that social intelligence is responsible for the emergence of folksonomies, and spend less time

on the analysis of the semantics embodied within the folksonomies. That we leave for another paper.

THE BENEFITS OF TAGS

The basic idea of user tags is extraordinarily straightforward: give a user a resource and ask him/her to assign keywords to them. No restriction or advice on the identity of permissible tags, or their number. Here is a description from Shirky (2005): "There is no fixed set of categories or officially approved choices. You can use words, acronyms, numbers, whatever makes sense to you, without regard for anyone else's needs, interests, or requirements." The point is that there is a very low entry cost to user tagging since it involves no training or instruction whatsoever, and very little work on the part of the users. Yet the benefits are great. For example on the social bookmarking service "del.icio.us", users mark up their favorite web sites with their chosen tags. Del.icio.us is a web site which requires a user account, and which acts in the first instance as a web based repository for each individual user's bookmarks for their favorite web sites. The web sites are indexed by URL and described with a textual description which is typically generated from the title in the web site. As a result, most bookmarks to the same URL will have the same descriptive title, but this is not necessarily the case because users are free to insert their own descriptions. In addition, users are free to annotate each bookmark with any number of single word tags. The user interface provides access to popular tags for a given URL, assuming that other users have tagged that URL. In addition, users can view other URLs annotated with a particular tag. Because the aggregated "tag use" of all users is available in various forms, users can derive value from each others behavior. For example popular tags for a given URL can influence a user who is also adding that URL to their bookmarks, because popular tags are, putatively, useful for other users. On the other hand, users can find new web sites by following links that were tagged with the same terms as the current one of interest. As pointed out in Udell (2004) the novel feature of services like del.icio.us is not their reliance on keywords in lieu of taxonomies for indexing -- that idea has been around for years. Instead, the novelty is the immediacy of the feedback from the community of users: "Feedback is immediate. As soon as you assign a tag to an item, you see the cluster of items carrying the same tag. If that's not what you expected, you're given incentive to change the tag or add another ... you can adapt to the group norm, keep your tag in a bid to influence the group norm, or both."

The "social" aspect of the system is that it fosters communities of interest in which groups of users can be identified with similar interests, facilitating knowledge discovery and sharing. But the benefits to indexing are that resources are grouped according to flexible category structures that are not imposed by authority. But how can this help in organizing resources?

One promised advantage of categorizing by tagging is that it facilitates classification by multiple aspects. The following example from Golder and Huberman (2005) illustrates the point nicely. In this example, formal category structures are likened to hierarchically organized "folders" in a computer file system, where files are stored in a single location on the file system.

"For example, consider a hypothetical researcher who downloads an article about cat species native to Africa. If the researcher wanted to organize all her downloaded articles in a hierarchy of folders, there are several hypothetical options, of which we consider four:

1.c:\articles\cats	all articles on cats
2. c:\articles\africa	all articles on Africa
3. c:\articles\africa\cats	all articles on African cats
4. c:\articles\cats\africa	all articles on cats from Africa

Each choice reflects a decision about the relative importance of each characteristic. Folder names and levels are in themselves informative, in that, like tags, they describe the information held within them. Folders like 1. and 2. make central the fact that the folders are about "cats" and "africa" respectively, but elide all information about the other category. 3. and 4. organize the files by both categories, but establish the first as primary or more salient, and the second as secondary or more specific. However, looking in 3. for a file in 4. will be fruitless, and so checking multiple locations becomes necessary."

The promise of tags is that they will eliminate the need to check multiple locations because they eliminate the need to "guess" which category to look in: you simply search for "Africa+cat". But this idyllic situation soon begins to look a little worse … suppose the researcher downloads some more articles, this time specifically about "cheetahs". Neglecting to tag them with the existing tag "cat", she decides to use the more specific tag "cheetah". But now the search for "Africa+cat" fails to find these important articles! So the user has to look elsewhere, possibly realizing that the third tag is also relevant. This needs sophisticated tools that can do some fancy computations over the set of existing tags, involving information extraction and reasoning techniques. How bad will this get with millions of resources and millions of tags? No one knows, but some

speculate the confusion will become so big, and the costs of retrieving useful information in such a landscape so prohibitive, that the whole enterprise will one day disappear! But let us not be so pessimistic.

In order for a system of this sort to be effective in fostering reuse and discovery, it seems that *some* agreement between at least *some* sub-sets of users must exist. That is, users must agree roughly on what "articles", "cats", and "cheetahs" mean, if they are going to annotate articles about cats under those tags. Of course users are free to tag crocodiles, large buildings and mathematical laws with the tag "cats", but this will soon make the system collapse. On the other hand tags like "Africa" will, it seems, be used with more heterogeneous content: it will probably include sites about African politics, weather, health, and so on, as well as animals. But this is perfectly natural in a system of classification which admits different levels of generality. So tags do appear to offer benefits if looked at as collective systems of classification. But this view is resisted by the most radical advocates. Shirky (2006) writes: *Here's what's radical about what del.icio.us protends: My vocabulary on del.icio.us folksonomy is personal, not vernacular — no one knows or needs to know which class I'm talking about when I tag something 'class', or that I use LOC to mean Library of Congress. This isn't the same as, say, the dictionary of thieves slang from the mid-18th c. because no one else needs to know my bookmark system, and I don't need to know anyone else's".* This is a radical opposition to the "collective intelligence" notion of folksonomy. But can it be sustained? I don't think it can, for the following reasons. Here is what's radical about what Shirky portends: if I want to, I can tag my bookmarks with any vocabulary I chose. But surely this is formally too powerful a system. Suppose I made up tags like: "hdfjkfb", "orjfkido", "hjfoå", "krlofpke", where somehow I learned what each tag referred to. This certainly ensures that no one else knows my system, and if everyone else does the same, I won't know theirs. But how useful would this be to anyone?? Clearly at this rather radical extreme the Shirky claim is without content. But maybe the example is too radical because the notion of vocabulary precludes the use of un-systematic tags. Let us think of a different example with some systematicity. So I make up a system that only I know, where all sites I judge to be interesting end with "xyz", technical sites begin with "krp", and so on. Everyone else can make up their own system, and no one knows anybody else's. But an immediate problem with this is that it is cheating … it smuggles in the more natural English vocabulary via the back door by simply equating each English term with an expression in the new "vocabulary system". Still, this is private knowledge so maybe that is O.K. for the example. So how would such a system work? Suppose I made up a number of such vocabularies, and got different users to adopt them. To the individual user, each system would be equivalent to the English terminology. But which vocabularies would make for a better del.icio.us? The experiment hardly needs to be done.

I think it is pretty clear that we **do** need to know **something** about everyone else's bookmark system! The success of a "social bookmarking system" *depends* on the fact that we do understand each others vocabularies (to some extent), and can extract value from that shared understanding.

There is plenty of evidence in the aggregate data to show that a coherent, collective, and largely shared system of classification emerges from the practice of communal tagging.

EMERGING PATTERNS

To help gather data we used a freely available web site, cloudalicious (http://cloudalicio.us/), which provides visualizations of the historical patterns in tag use for a given URL.

Figure 2 shows a visualization of the way in which the most popular tags emerged for the New York Times web site.

There are a number of interesting observations that can be made about this graphic.

First, there is a hint of the "power law curve" (Hyde, 2005) apparent in the curves. The idea is that there are a few tags which are used very often and very many tags which are used less often. In this example the use of the tag "news" dominates and the rest of the tags are much less common. This particular example is a bit atypical because the most frequent tag dominates more strongly than is typically the case, with the least popular tags being a little too close together at the bottom. But the most striking observation is how much agreement there is in the use of the tags. After a brief unsettled period at the beginning of the history, the pattern pretty much stabilizes and the dominance of the most popular tags is never challenged. Scott Golder and Bernardo Huberman at HP Labs performed an extensive study of tagging behavior and come to the same conclusion about the amazing stability of tag use. But they make two additional points of interest. First, they show that tags tend to stabilize after just 100 bookmarks have been assigned to the URL. Thus even relatively unpopular sites end up with stable tag clouds. Secondly they argue that "imitation" made possible by the user interface can't be the whole explanation of the stability of tags since the less popular tags, which are not shown as suggestions through the interface, nevertheless display the same stable patterns over time (Golder and Huberman, 2005). This is an important point because it shows that individual users are independently assigning the same tags to the same resources with sufficient frequency to become observable and stable in the aggregate data.

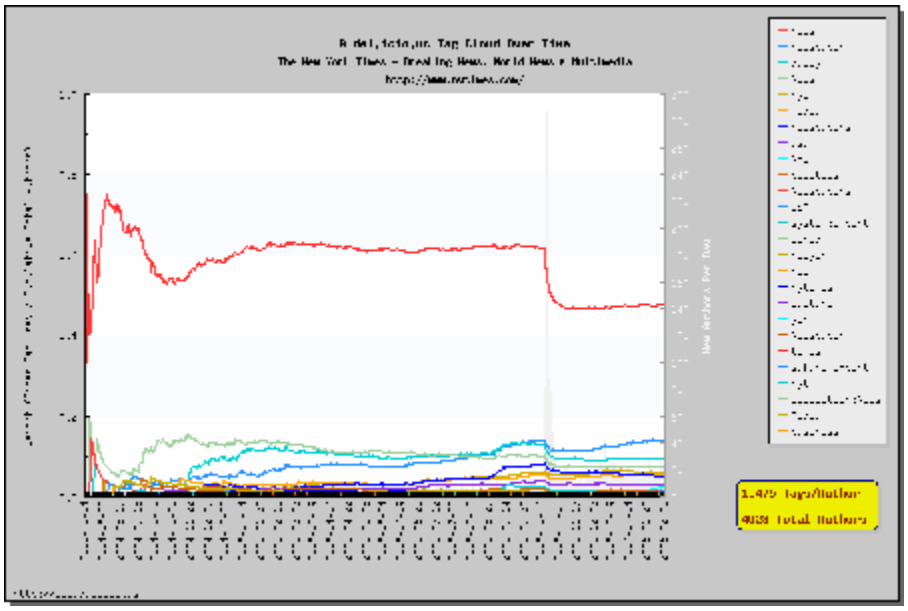

Figure 2. Historical view of tag cloud for The New York Times

This observation has important consequences for the power law curve observed with tags. Tim Vanderwall, the inventor of the term "folksonomy", attempts to explain the power law curve with a process in which people are brought together into clusters who share common vocabularies because it helps them identify the resources (Vanderwal, 2005). Since new users typically have access to the most popular tags already used for that site, it is likely that they will chose at least some of those existing tags. But this sets up a positive feedback loop. If the popular tags are picked by new users they will become more popular still, which will influence even more new users to pick that tag, and so on. But this "social" explanation can't be the whole story. Consider the following experiment: Suppose I am a really rich guy who wants to influence tags on del.icio.us. So I pay 10000 people to tag resources according to my schema. Suppose I wanted the most popular tag for each URL to be some sort of emotional evaluation like "cool", "interesting", "awful", and the like. Would these stick? My feeling is, NO. The prediction is that "subjective" tags of this sort won't make it because there is too much individual variability in the emotional reaction to a site … and this reaction is hard to coerce. On the other hand the tags which do survive the "popularity contest" are those about which, contrary to claims we have discussed, there is not much disagreement and individual difference. Put another way, even though I might not have thought to label "The New York Times" with "News" on a particular occasion, I certainly would not argue that it should not be labeled with "News". These tags, then, could be the collective emergent categories observed in folksonomies.

Unfortunately the discovery of collective categories is not as straightforward as simply extracting the most popular tags. Consider table 1 which lists the four most popular tags for some of the top 50 most popular sites from delicious (as tabulated by the web service at http://populicio.us/):

It is self evident that the popular tags form a heterogeneous collection. Some are clear category labels that would feel at home in a formal taxonomy (e.g. "News", "Movies", "Music"). Some, like "Daily" and "Recommendation" appear to describe resources with a particular property which is nevertheless fixed and user independent. Others like "Fun" and "Geek" describe more personal properties that depend on individual interpretation. Finally there are proper names like "UK" and "NYC".

Clearly then, there is some consensual pattern emerging, but it is not a straightforward one. In the next sub section we present a novel analysis that sheds some light on the underlying mechanisms at work in creating these agreements, and therefore helps expose the meanings behind the folksonomic terms.

Site	Slashdot	Flickr	Pandora	Digg	BBC News	New York Times	Internet Movie Database	PocketMod	Boing Boing
Tags	News, Technology, Geek, Daily	Photos, Flickr, Photography, Sharing	Music, Radio, Recommendation, MP3	News, Technology, Blog, Daily	News, BBC, UK, Daily	News, Newspaper, Daily, NYC	Movies, Reference, Film, Database	Productivity, GTD, Organization, Lifehacks	Blog, News, Daily, Fun

Table 1. A list (in descending order) of the four most popular tags for the corresponding site. Each site is in the top 50 most popular sites on del.icio.us

The Ontology of Folksonomies

In case our conceptual bias is not self evident, we state it here to avoid any confusion. Our hypothesis is that mental architecture fundamentally shapes our perceptions and organization of the world in which we live. Further, essential aspects of the mental architecture are fixed and therefore shared by all humans, which is what makes communication and shared understanding possible. The overlap is not perfect. I say "Library of Congress", but Clay Shirky wants to say "LOC". This is more like noise than disagreement in our view. But pity the poor soul who calls it “the square root of negative 2″! At the right level of abstraction we think alike. The mind creates categories, because that is what minds do. The mental architecture enforces the range of possible ontologies and taxonomies that we can bring to bear on the understanding of our universe. All humans share fundamental aspects of mental architecture and therefore properties of possible taxonomies. Folksonomies provide a fantastic window into the workshop of the mental taxonomist: the emergence of "social intelligence" is a reflection of the cognitive architecture we all share. The remaining problem is to extract regularities from folksonomies at the most useful and explanatory level of abstraction.

We argue that the linguistic properties of tags provide the best clue to their underlying nature. There is a rich tradition in modern linguistics to connect aspects of semantic structure with their syntactic realizations (e.g. Jackendoff, 1983; Pinker, 1989; Levine, 1993). We report a study in which we compare folksonomies with the category structure of Yahoo directory and DMOZ (Veres, 2006). To a first approximation, we note the common generalization that nominals tend to represent categories and adjectives describe properties, and find that one major difference is that only folksonomies make extensive use of adjectives. But we also find that the category "nominal" is too broad for a useful comparison and we develop a novel approach to analyzing folksonomy categories. The analysis is based on the work of linguist Anna Wierzbicka who described several different kinds of categories that people employ in their cognitive organization of the world. She notes that in the biological world taxonomic categories are common, so that "cats" and "dogs" are taxonomic sub classes of "animal". On the other hand while artifacts are often described (erroneously) as taxonomies, they are in fact grouped by other principles. A category like "furniture" is formed, on this view, because its members are often experienced together in a common location and serving a common function. The members comprise a very loose and heterogeneous collection which might include tables, chairs, lamps, ashtrays, stereo systems, televisions, and so on. Thus a chair is not a-kind-of furniture. Wierzbicka proposes a number of such categories and, importantly, claims that the categories can be distinguished by the syntactic properties of their names. Thus for example, *furniture* is syntactically *singularia-tantum*, meaning it does not appear as a plural. Since the semantic categories can be distinguished by their syntactic properties, it is possible to define a set of syntactic frames that can be used to decide which category a particular word represents, and therefore identify the sort of category that each folksonomy term represents. Using these techniques we have been able to define a layer of abstraction on the aggregate, popular tags, which can organize them in useful ontological structures. The key is to organize nominals according to their abstract categories in a way that maintains certain relationships between the abstract types. For example if a web site for chairs was tagged with "chair" and "furniture", we would argue that the collective concept "furniture" should subsume the taxonomic "chair". (Note *chair* has a completely different syntax from *furniture*: "one chair" vs. "one furniture", "three chairs" vs. "three furnitures", etc.). Using these principles we can assemble the confused array of terms in the folksonomy into a coherent ontology. We take this as *existence proof* that folksonomies display patterns of formal classification, driven by sub conscious cognitive processes. Our current research involves the definition of algorithms that can structure folksonomies according to abstract linguistic definitions in a fully automatic fashion.

CONCLUSION

In this paper we have presented evidence that social tagging results in folksonomies filled with rich and exploitable meaning. The arguments presented have a theoretical and practical significance.

Theoretically we address the connection between linguistic and cognitive facts to the emergence of social intelligence. We argue that the mental architecture constrains directly the nature of the emergent classification schemas which, contrary to popular belief, display great deal of abstract structure.

But it also has great practical implications since it promises to unite two revolutionary new technologies: the emerging Web2.0 and the Semantic Web. The ability to combine the flexible and dynamic nature of the emerging Web2.0 technologies with the possibilities afforded by the semantically rich languages of the Semantic Web offers tremendous promise for both. As a starting example, the foreshadowed integration of resources in different Web2.0 services will be greatly facilitated by our work.

ACKNOWLEDGMENTS

This work was partly sponsored by the Norwegian Research Council through the WISEMOD project, number 160126V30 in the IKT-2010 program.

REFERENCES

1. Baron-Cohen, S. (1997). *Mindblindness*. Bradford Books, MIT Press, Boston, MA.
2. Davis, I. (2005). Why *Tagging Is Expensive*. http://silkworm.talis.com/blog/archives/2005/09/why_tagging_is.html
3. Golder, S., and Huberman, B. A. (2005). *The Structure of Collaborative Tagging Systems*, Citebase:http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0508082
4. Hyde, B. (2005) *Tagging Power Law.* http://enthusiasm.cozy.org/archives/2005/01/tagging-powerlaw
5. Jackendoff, R. S. (1983). *Semantics and Cognition*. MIT Press, Cambridge, MA.
6. Levine, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
7. Pinker, S. (1989). *Learnability and Cognition*. MIT Press, Cambridge, MA.
8. Speroni, P. (2005) *Tagclouds and cultural changes*. http://blog.pietrosperoni.it/2005/05/28/tagclouds-and-cultural-changes/
9. Speroni, P. (2004). *Hierarchical Delicious Free Mind Map*. http://blog.pietrosperoni.it/2004/09/06/hierarchical-delicious-free-mind-map/
10. Udell, Jon. (2004) *Collaborative knowledge gardening.* InfoWorld. August 20, 2004. http://www.infoworld.com/article/04/08/20/34OPstrategic_1.html
11. Vanderwal, T. (2005) *Explaining and Showing Broad and Narrow Folksonomies.* http://www.vanderwal.net/random/entrysel.php?blog=1635
12. Veres, C. (2006) The Language of Folksonomies: What Tags reveal about user classification. *Proceedings*, NLDB 2006, Springer-Verlag.
13. Wierzbicka, A. (1984). Apples are not a "kind of fruit": the semantics of human categorization. *American Ethnologist*, 11, 313-328.