Semantic Halo for Collaboration Tagging Systems

Alan Dix[1], Stefano Levialdi[2], Alessio Malizia[2]*

[1] Lancaster University, Lancaster, LA1 4YR, UK
[2] University "La Sapience" of Rome, Via Samaria 113,
00198, Rome, Italy
alan@hcibook.com, {levialdi,malizia}@di.uniroma1.it

Abstract. Collaborative tagging systems allow many users to add keywords (tags) to community-shared data items. Recently, collaborative tagging systems, also known as folksonomies, are growing on the web allowing people to annotate content, and then query by submitting keywords or tags. In this paper we define a *Semantic Halo*, as a set of additional information that can be provided from tagging systems to end users when retrieving user-relevant documents. We analyze the semantic aspects of tagging and provide an algorithm for computing the *Semantic Halo* of tags. Finally, we show some preliminary results that demonstrate the effectiveness of our approach.

1 Introduction

Collaborative tagging describes the mechanism by which many users add metadata in the form of keywords to community-shared content. Recently, collaborative tagging has become popular on the web, in fact many web sites allow users to tag bookmarks, photographs and other document types. Document repositories or digital libraries often support documents' organization by assigned keywords. By contrast, usually such classification is either performed by an authority, such as a librarian, or else emerges from the material supplied by the authors of the documents [1] [4]. Conversely, collaborative tagging is the method of allowing anyone (users) to link keywords or tags to content, at pleasure. Collaborative tagging systems are an alternative mechanism to the semantic web approach where experts build ontologies [2] with predetermined relationships among keywords. This later approach, usually, requires domain-field experts and a community agreeing on most of the experts choices; while in collaborative tagging, keyword indexing grows as a natural process. Collaborative tagging systems are also known as "folksonomy", which stands for "folk taxonomy", since by adding metadata to documents a community builds a personalized taxonomy. Many examples of these tools are present on the web, such as: *Del.icio.us* that permits collaborative tagging of shared website bookmarks, *Snipit*, which is also able to bookmark sections of web pages and *CiteULike* or *Connotea* that allow

* Corresponding author.

the same for references to academic publications. Some services allow users to tag, but only on content they own, for example, *Flickr* for photographs and *Technorati* for weblogs.

From a user perspective, navigating a tagging system is similar to performing keyword-based searches users provide salient, descriptive terms in order to retrieve a set of related items. Our approach could be thought as a semantic tagging expansion of terms for augmenting querying experience in folksonomy systems. In fact, in order to get broad coverage so that many possible queries can be formulated, the meanings for individual tags (terms) are expanded by a tags cloud called *Semantic Halo*, which retrieves additional related information to the submitted tags. For instance, if the user is searching for documents annotated with the tag "University", it could be useful to retrieve also documents including related concepts (thus tagged with different but semantically related terms) such as documents about "colleges" or "education". These last two terms are respectively considered by the system as specification, and generalization of the submitted tag. They will be included, by employing the *Semantic Halo* approach, in the submitted query and thus in the list of documents retrieved by the system.

In this paper we define a *Semantic Halo*, as a set of additional information that can be provided from tagging systems to end users when retrieving relevant documents. We analyze the semantic aspects of tagging and provide an algorithm for extracting the *Semantic Halo* of tags. Moreover, we show some preliminary results that demonstrate the effectiveness of our approach.

2 Related Works

Recently, collaborative tagging has grown in popularity on the web, and researchers both at academic and industrial level are starting to produce papers on this subject. In [1], Golder *et al.* explore the structure of collaborative tagging systems as well as their dynamical aspects. Particularly, they show regularities in user activity, tag frequencies, kinds of tags used, popularity in bookmarking and a notable stability in the relative proportions of tags within a given url (related to the *Delicious* system). Moreover they present a dynamical model of collaborative tagging that predicts these stable patterns and relates them to imitation and shared knowledge. Differently from our approach their system is related to the frequency of tagging for a given url, while we explore information for a given tag or term. The creation of metadata has generally been approached in two ways: professional creation and author creation. In libraries and other organizations, creating metadata, primarily in the form of catalog records, it has traditionally been the domain of dedicated professionals working with complex, detailed rule sets and vocabularies. In [4], Mathes observes that the primary problem with this approach is scalability and its impracticality for the vast amounts of content being produced and used, especially on the World Wide Web. In fact systems developed around professional cataloging are usually too complicated to use for anyone without specific training and understanding. The Mathes paper examines the folksonomy approach: user-created metadata, where users of the documents and media, create metadata for their own individual use also shared throughout a community. It should be noticed that an important aspect of a folksonomy is that it is comprised of terms in a flat namespace: that is, there is no hierarchy, and no directly specified parent-child or sibling relationships between these terms. In his paper, Mathes illustrates further research areas like: quantitative task analysis, qualitative user analysis, applicability to other systems.

We propose instead a *Semantic Halo* approach where relationships among tags include things like broader, narrower, as well as related terms, with respect to a specific tag or term. In [5], it is correctly stated that tag clouds are becoming more and more popular. Tag clouds and tag sets are a different kind of objects. As a tag set, the author means a set of tags. With no order, either a tag is part of the set, or it is not. The tags that one user deploys to bookmark a single url in *del.icio.us* is a tag set (or a tagset). Tag clouds (or tagclouds) are a multi-set of tags. That is, a set of tags where each tag can appear with multiplicity higher than one. Examples of the first type of tools are *Flickr*, *43things*, *consuMating*, *tagsurf*; an example of the second is the tagged version of the *BBC* web site contents. In all these cases a tag set is used, where perhaps a tag cloud would be more appropriate. Some of the differences between a tag cloud and a tag set were explained in [6]: ”Explaining and Showing Broad and Narrow Folksonomies”. In our approach, we decided to use tag clouds as we perform reasoning and frequency related computation, not simply tag set exploration. In [5], another relevant factor is explored, that inspired us in the definition of our *Semantic Halo*; the time. Time is an important factor in considering collaborative tagging systems, in fact definitions and relationships among tags could vary over time. A clear example is given in [5], where tagging of the paper ”Power Laws, Weblogs, and Inequality” by Clay Shirky in the *Delicious* community suddenly changed over time. When it came out, the term ”long tail” was not used; long tail in this article is referred to long tail of weblogs. Long tails were always present, they were just not culturally recognized as such. On the October 2004 issue of the article from Wired, ”The Long Tail” came out. The article was an immediate hit, and on the same day in which the first person bookmarked the article 21 other persons bookmarked it too. The link appeared on *Delicious* popular, and a huge number of people read it, and bookmarked it. This article changed the way people perceived the previous article from Clay Shirky. Today the only tags more common than ”longtail” are: powerlaw, blogs, blog, blogging, web, network socialsoftware, shirky.

3 Semantic aspects of tagging

In this section we describe a list of considerations about semantic aspects of tagging that lead us to develop the *Semantic Halo* approach.

By providing meanings at different levels of generality, the system can produce alternative queries for selection by the user that span a wide range of possible interpretations, also considering variation over time, as we show in the next paragraph. Reflecting on the cognitive aspect of hierarchy and categorization, the "basic level" problem is that related terms that describe an item vary along a continuum of specificity, ranging from very general to very specific; the problem lies in the fact that different persons may consider terms at different levels of specificity to be most useful or appropriate for describing the item in question. We address this problem by including in, our *Semantic Halo* , an *abstraction* feature. This *abstraction* feature retrieves related tags by increasing or decreasing the level of generalization. We retrieve terms that have an increasing level of generalization and thus are generalization of the given tag among the tagging community. Moreover we also retrieve terms that have a decreasing level of generalization and are thus considered specialization of the given tag among the tagging community. For the purposes of tagging systems, however, conflicting basic levels can prove disastrous, as documents tagged "perl" and "javascript" may be too specific for some users, while a document tagged "programming" may be too general for others, as stated in [1]. Tagging is fundamentally about sense making. The underlying factor behind this variation may be that basic levels vary in specificity to the degree that such specificity makes a difference in the lives of the individual. Like variation in expertise, variations in other social or cultural categories likely yield variations in basic levels. Thus we think that retrieving generalization and specification related tags could help users in finding the right item they look for in the tagging system, being robust against socio/cultural variations among tags.

Collective tagging, then, has the capacity to exalt the problems associated with the fuzziness of linguistic and cognitive boundaries. As all tagging community contributions collectively produce a wider classification system, that system consists of personal classifications as well as those that are widely agreed upon. Thus, both tagging systems and taxonomies are affected by many problems that exist as a result of the necessarily inexact, but instinctive and evolving process of creating semantic relations between words and their referents. Three of these problems are polysemy, synonymy, and basic level variation.

A polysemous word is one that has many ("poly") related senses ("semy"). In practice, polysemy alters query results by returning related, but potentially irrelevant, items. Superficially, polysemy is similar to homonymy, where a word has multiple, unrelated meanings. Synonymy, or multiple words having the same or closely related meanings, presents a greater problem for tagging systems because inconsistency among the terms used in tagging can make it very difficult for one to be sure that all the relevant items have been found. This problem is compounded in a collaborative system, where all taggers either need to widely agree on a convention, or else accept that they must issue multiple or more complex queries to cover many possibilities. Synonymy is a significant problem because it is impossible to know how many items "out there" one would have liked one's query to have retrieved, but didn't. Basic level variations like plurals and parts of speech and spelling can also stymie a tagging system. We deal with these problems by including in our *Semantic Halo* contextual information. We retrieve multiple contexts associated with given tags (terms) and in this way, by measuring correlation among contexts we can separate them (meanings) among different community tagging senses. In this way we deal with polysemy and if the opposite is true (contexts are very similar), we can deal with synonyms and thus retrieve very similar contexts. By employing this approach we notice that we can also include basic level variations as shown in the next paragraph. We must say that there are still some errors in retrieving *Semantic Halo* information, but early results are very encouraging. Moreover, dealing with contexts in our case is also related to include variations over time in community tagging. In fact in our *Semantic Halo* we return the contexts ordered by time, and thus we can also capture meanings' variation over time, as we will show later in next paragraph.

4 Extracting the Semantic Halo

As discussed earlier, a folksonomy represents a fundamental shift in that it is derived not from professionals or content creators, but from the users of information and documents. In this way, it directly reflects their choices in diction, terminology, and precision. There is no significant cost for a user or for the system to add new terms to the folksonomy. The problem is that while the disparate user vocabularies and terms enable some very interesting browsing and finding, the sheer multiplicity of terms and vocabularies may overwhelm the content with noisy metadata that is not useful or relevant to a user. Furthermore, the cost for users of the system in terms of time and effort is far lower than systems that rely on complex hierarchal classification and categorization schemes. In addition to this structural difference, the context of use in these systems is not just one of personal organization, but of communication and sharing. The nearly instant feedback in these systems leads to a communicative nature of tag use.

These considerations drove us in thinking that a *Semantic Halo* will help users in augmenting querying results for matching as close as possible to the user's needs. We define our *Semantic Halo* as a set of search results for a given tag made by a set of four features, we named it **4A**:

- **Aggregation**. It contains all the tags (terms) linked or related to the given tag.
- **Abstraction**. It is similar to aggregation but related to a direction (increasing and decreasing), thus it contains two subsets: *Generalization*, tags increasing abstraction with respect to the given tag, and *Specialization*, tags decreasing abstraction with respect to the given tag.
- **Ambience**. It is the context for a given tag. Thus it includes all the possible tags appearing in the same context, and that will be useful for augmenting or refining the user query. This set will be built from a basic *Context* set, as clarified later.
- **Age**. It is a list of ordered contexts, namely an ordering of the Ambience feature elements over time. This will help in retrieving tags ordered by meanings given to them during time.

More formally we define these **4A** features as sets:

Given a tag t, let Doc be the set of documents being tagged, consider $d \in Doc$, and write $tags(d)$ as the set of tags for d in Doc, now set: $n_j = co\_occur(t, t_j) = card\{d \in Doc | t \in tags(d)\ and\ t_j \in tags(d)\}$, and $C_t = \bigcup\{tags(d) | t \in tags(d)\} - \{t\} = \{t' | t_i \neq t\ and\ co\_occur(t, t') > 0\}$.

Finally, we define the average instance frequency of $f_t = \frac{1}{m} \sum_{i=1\ldots m} N_i$.

```
Let
$Aggregation = Abstraction = Genralization = Specialization = \emptyset$
$Ambience = Context = Age = \emptyset$ .
//all sets are empty at the beginning

For each $t_i \in C_t$ extract
$C_{t_i} = \{t_{i1}, \ldots, t_{in}\}$ //the set of $t_i$ co-occurrence tags
Let $N_{t_i}$ //the number of co-occurrence tags instances

let $t \in C_{t_i}$ and let $j \mid t_{ij} = t$ then
    if ($n_i > f_t$) and ($n_{t_{ij}} > f_{t_i}$) then $Context = Context \bigcup t_i$
    if ($n_i > f_t$) and ($n_{t_{ij}} \leq f_{t_i}$) then
     $Generalization = Generalization \bigcup t_i$
    if ($n_i \leq f_t$) and ($n_{t_{ij}} > f_{t_i}$) then
     $Specialization = Specialization \bigcup t_i$
    if ($n_i \leq f_t$) and ($n_{t_{ij}} \leq f_{t_i}$) then
     $Aggregation = Aggregation \bigcup t_i$

Set $Abstraction = Generalization \bigcup Specialization$
```

After this processing phase, and before building the Age set, we partition the $Context$ set in a new set made of $Context$ subsets called $Ambience \subseteq P(Context)$, which is contained in the power set of $Context$.

We explore the $Context$ set, and for each $t_h, t_k \in Context$ we compute N_{t_h}, N_{t_k} and their respective frequency; if it is higher than f they are part of the same context and $Ambience = Ambience \bigcup\{t_h, t_k\}$; else $Ambience = Ambience \bigcup\{t_h\} \bigcup\{t_k\}$. By building these subsets, we have subsets of tags that represent the same context (meaning) thus could be considered as synonyms, and other distinct subsets represent separate meanings, thus show polysemic meanings of the same tag. Finally, we build the Age feature as an ordered sequence, by ordering the $Ambience$ subsets by age. Ordering by age means that for each subset contained in $Ambience$, we take the date of the most recent submitted tag.

In order to show the results of our approach, we tested it among the *Delicious* community. Delicious is a social bookmarks manager on the web. Users submit their links to a website, adding some descriptive text and keywords, and *Delicious* aggregates their post with everyone else's submissions allowing users to share their posts. We implemented our algorithm for extracting the *Semantic Halo* using their programming APIs (Application Programming Interfaces) and thus obtaining results while community users where tagging. We present and comment our early findings, but we notice that the Delicious community is very large (now

Delicious is part of Yahoo) and moreover it is a very active tagging community and this results in a quite complex but effective test. We tested our approach with different tags we believe are interesting for demonstrating early results, and we show the features extracted from our algorithm. We think that a simple interface could be developed which presents to end users not only the retrieved bookmarks by their submitted tags, but the four features provided by our algorithm for enhancing the search/browsing experience.

Given the tag "university", which is quite general, our algorithm searched over Delicious for related tags and retrieved:

- **Ambience** = { open, { learning, University } }
- **Abstraction** = { online, education} $\bigcup$ {colleges, high, degree, distance, Commons }
- **Age** = ((learning,University), (open))
- **Aggregation** = { soccer, gradschool, corps, indoor, course, masters, research_institute, cites, cincinatti, peace, demographic, content, courses, innovators, urban, tournament, entrepreneurship, liverpool, york, community-college, schools, Illinois, abroad, Content, latino, Course, complexity, planning, Initiative, academiclibrary, enterprise, semantic_web, Education, grad, scholarship, teaching, college, school }

We can observe that the **Ambience** set is composed of two subsets associated with two different contexts and thus meanings of "university" tag. Interestingly, we can see that we can solve also basic level variations since the tag "University" with the capital "U" is strongly associated with the "university" tag and also together with "learning" could be considered as a synonym; while "open" is also strictly related but indicates a different meaning thus coping with polysemy. The first part of the **Abstraction** set is related to the generalization of the given tag, while the second part is specialization, thus providing a partition of the related tags in increasing and decreasing abstraction. The **Age** sequence is the ordered set of contexts (meanings) with respect to last updates. The **Aggregation** set lists all the related tags, and even if there are unwanted tags the majority is clearly related.

Another interesting example occurs with the tag "math":

- **Ambience** = { Math }
- **Abstraction** = { programming} $\bigcup$ {fractal, algorithmic, formal }
- **Age** = ((Math))
- **Aggregation** = { genprog, engineering, poems, nerd, books, statistics, poetry, children, application, stories, finance, latex, reading, data, articles, kids, Goose, Mother, book, crafts, sparlings, research_result, Programming }

We can observe that again the basic level variation is solved since we identify "Math" as strictly closed to "math". But there are two other interesting findings to be observed. First of all the community for now is still quite biased by being accessed mainly by computer-related people. We can see that "programming" is a generalization of "math" in this environment and also tags like "genprog" and "latex" are aggregated to "math". But another interesting phenomenon is that of finding "kids" and "children" aggregated by "math" with tags like "Mother" and "Goose" which are related to maths books for children. This is very interesting and we think very helpful to users searching for links on the math subject. Moreover specializations retrieved for "math" tag are very meaningful like "algorithmic" and "formal".

We presented these two, we believe relevant, results for this very first version of our approach. We have many of these examples but they are somehow related to the programming and computer area, even if we think that since this community is growing very quickly we can have, soon, many interesting examples of general tags and concepts for testing our *Semantic Halo* approach.

5 Conclusions and Future Works

We believe that providing the *Semantic Halo* as result of a query for a given tag will strongly help users in finding desired items in community tagging systems. In fact, instead of retrieving simply a related list of tags, as it happens, mainly, in all available tagging systems, we present to users four classes of grouped tags that are not only related to the submitted one, but also provide useful information for avoiding typical problems (synonyms, polysemy, basic level variations) of community-based tagging systems.

We are planning to use our *Semantic Halo* for conducting usability experiments among users to show its validity in augmenting seraching/browsing. We will explore different tagging systems and folksonomies, not only for validating our approach but also for investigating if the *Semantic Halo* can be employed for managing and exploring tagging communities having cultural and social bias.

References

1. S. A. Golder and B. A. Huberman, "The Structure of Collaborative Tagging Systems". To appear in the *Journal of Information Science* (2006)
2. C. Shirky, "Ontology is Overrated: Categories, Links and Tags". *www.shirky.com/writings/ontology_overrated.html.* (2005)
3. *www.google.com/gmail*
4. A. Mathes, "Folksonomies Cooperative Classification and Communication through Shared Metadata". *Computer Mediated Communication - LIS590CMC Graduate School of Library and Information Science, University of Illinois Urbana-Champaign.* (December 2004) *www.adammathes.com/academic/computer-mediatedcommunication/ folksonomies.html*
5. Speroni, P. "On Tag Clouds, Metric, Tag Sets and Power Laws". May 2005. *blog.pietrosperoni.it/2005/05/25/tag-clouds-metric/*
6. Explaining and Showing Broad and Narrow Folksonomies. 2005. *www.vanderwal.net/random/entrysel.php?blog=1635*