Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems

Paul Heymann* and Hector Garcia-Molina†
Computer Science Department, Stanford University
Stanford, CA 94305
(Dated: April 24, 2006)

Collaborative tagging systems—systems where many casual users annotate objects with free-form strings (tags) of their choosing—have recently emerged as a powerful way to label and organize large collections of data. During our recent investigation into these types of systems, we discovered a simple but remarkably effective algorithm for converting a large corpus of tags annotating objects in a tagging system into a navigable hierarchical taxonomy of tags. We first discuss the algorithm and then present a preliminary model to explain why it is so effective in these types of systems.

I. INTRODUCTION

One of the lasting problems for CSCW in large scale organizations has been determining an effective way to not only store, but to allow users to annotate relevant data for future retrieval. As the size of our information systems have expanded, there has been a gradual trend from centrally organized systems based on controlled vocabularies (i.e., a library model) to chaotic, ad-hoc distributed systems with many cooperating participants. This spectrum represents a trade-off that must be made between how much effort is required to make a single annotation and how much of the data is annotated.

Perhaps the furthest points on this spectrum are collaborative tagging systems where users cooperate to label objects in large scale systems (the web, large media collections) with *tags*—free-form strings of their choosing. Popular tagging systems which have been studied in depth include Delicious [11], Flickr [15], and Connotea [14]. These systems allow any metadata to be associated with a given object, in contrast to stricter systems like libraries where a book will have exactly one proper call number based on content. As a result, users of tagging systems can quickly label (*tag*) large numbers of objects, but these labels are much less informative—tags tell us little more than the free-form string that they represent.

This lack of information in the tags reduces the ease of navigation in these systems. Currently, users can browse the objects in a tagging system using three main views:

1. A list of all objects which are tagged with a given tag (or possibly a combination of two or more tags).
2. A list of the most popular tags in the system.
3. A list of tags which have a high degree of overlap with a tag the user is currently investigating.

However, these limitations make it difficult to find broader or narrower tags which may better represent the user's current interests.

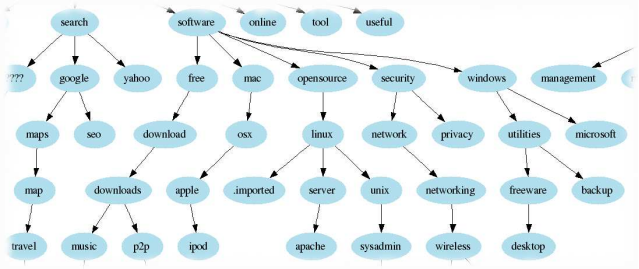


FIG. 1: A portion of the tag hierarchy generated from the Delicious dataset with a similarity threshold of 0.099 from tags occurring more than 400 times.

We have designed an algorithm which helps to solve this problem by automatically building a hierarchy of tags from the data in a tagging system. See Figure 1 for an example of a hierarchy produced from an unstructured set of tags found in the Delicious social bookmarking system. The algorithm leverages notions of similarity and generality that are implicitly present in the data generated by users as they annotate objects. Our algorithm brings tagging systems closer to the best of both worlds—such systems can leverage the contributions of huge numbers of casual users (Delicious claims 300,000 users [23]) while allowing users to use natural browsing conventions from more controlled and expensive systems. Taken together, we believe that the combination of large numbers of cooperating users and detailed analysis of the resulting data provides a compelling CSCW view for organizing the web and other very large scale information systems.

A. Related Work

The CSCW community has previously studied metadata annotations, how best to categorize these annotations, and systems for communal creation of knowledge around objects in large document systems. Annotations have been explored for large document and media collections, especially with respect to collaborative writing (for example, [3] and [20]). Recent work has found that

*Electronic address: heymann@cs.stanford.edu; Supported by an NSF GRFP Fellowship.; URL: http://i.stanford.edu/~heymann/
†Electronic address: hector@cs.stanford.edu

metadata in document collections has interesting social and contextual features (Marshall and Brush analyze differences between public and private annotations in [16], and Hinrichs *et al.* try to leverage context in metadata creation in [12]). Especially interesting has been work on how best to commonly create shared abstractions out of individual users' data which go above and beyond each user's contribution (see for example the work of Wu *et al.* on hierarchy metadata combination in [21] and [22]).

Collaborative tagging has put a new twist on these, and other, old CSCW problems. Tagging is similar to older work on adding keyword metadata to documents in document collections, though tagging systems are often substantially larger (even small scale tagging systems often have thousands of users),[1] less structured, and more egalitarian because users (rather than solely the creators of the content themselves) can often annotate any object in the system. Interestingly, the slew of related tags which appear as a result of this lack of structure are reminiscent of solutions like multiple synonymous indexes proposed for the "vocabulary problem"—where users often disagreed on the appropriate term to describe some concept or type of information—considered by the CSCW community at least two decades ago (see [5, 9, 10]). Finally, while traditionally CSCW has focused on how to elicit explicit contribution of community members (for example, [1]), it is becoming increasingly common in new CSCW systems (e.g., tagging) for members to make implicit contributions by acting on their own personal goals. This creation of knowledge in an indirect way is similar to, but slightly different from, the sort of tacit knowledge transfer discussed in [7] (though that work focuses on the problem of transferring difficult to explain knowledge through indirect means, rather than the general problem of indirect knowledge transfer).

Beyond these factors, taxonomy generation from unstructured data is a compelling problem in its own right. Some work has been done on taxonomy generation for large blocks of unstructured text, especially using clustering (see [6] for an example), but naive clustering approaches seem to fail for tag hierarchy generation, and this seems to be due to the structure of the data itself.[2]

B. Preliminaries

We assume that a minimal tagging system consists of objects $(o_1, o_2, ...)$, users $(u_1, u_2, ...)$, and tags $(t_1, t_2, ...)$. The data in a tagging system consists of annotations of objects $(a_1, a_2, ...)$, each of which contains one user (u_i), one object (o_j), and one or more tags $(t_k, ...)$. An annotation may also include other information in practice, like the date of the annotation or a personal note added by the user. We assume that no information is known about the relationships between the tags other than what can be implicitly derived from the objects which they annotate.[3]

Tags are aggregated into *tag vectors*, for which the index $v_{t_l}[o_m]$ is equal to the number of times that the tag t_l annotates the object o_m. We calculate the *similarity* between tags using the cosine similarity between tag vectors, although other similarity metrics may be acceptable as well. The *tag similarity graph* for a given dataset is an unweighted graph where each tag is represented by a vertex, and two vertices are connected by an edge if the similarity of the nodes they represent is above some set threshold.[4]

Algorithm 1 An extensible greedy algorithm for hierarchical taxonomy generation from social tagging systems using graph centrality in a similarity graph of tags.

Require: $L_{generality}$ is a list of tags $t_i, ...t_j$ in descending order of their centrality in the similarity graph.
Require: Several functions are assumed: $s(t_i, t_j)$ computes the similarity (using cosine similarity, for example) between t_i and t_j. $getVertices(G)$ returns all vertices in the given graph, G.
Require: *taxThreshold* is a parameter for the threshold at which a tag becomes a child of a related parent rather than of the root.

$G_{taxonomy} \leftarrow \langle \emptyset, root \rangle$
for $i = 1...|L_{generality}|$ **do**
  $t_i \leftarrow L_{generality}[i]$
  $maxCandidateVal \leftarrow 0$
  for all $t_j \in getVertices(G_{taxonomy})$ **do**
    if $s(t_i, t_j) > maxCandidateVal$ **then**
      $maxCandidateVal \leftarrow s(t_i, t_j)$
      $maxCandidate \leftarrow t_j$
    end if
  end for
  if $maxCandidateVal > taxThreshold$ **then**
    $G_{taxonomy} \leftarrow G_{taxonomy} \cup \langle maxCandidate, t_i \rangle$
  else
    $G_{taxonomy} \leftarrow G_{taxonomy} \cup \langle root, t_i \rangle$
  end if
end for

[1] Leveraging increasing numbers of casual contributors is a trend in CSCW systems (e.g., the Wikipedia). Recent work like [17] has tried to explain why people contribute to these sorts of systems.

[2] Similarity between parents and their children in a reasonable hierarchy does not seem to be sufficiently great for purely similarity based hierarchical clustering to produce useful results.

[3] Other systems for creating consensus hierarchies in CSCW systems, like the one discussed by Wu *et al.* in [21] and later in [22], rely on some degree of structure in per user tags.

[4] For tagging systems, this threshold is fairly obvious—there is a huge dropoff in similarity between unrelated tags. However, we intend to extend our representation to a weighted graph in the future.

II. ALGORITHM: EXTENSIBLE GREEDY MOST GENERAL FIRST

After having limited success producing hierarchical taxonomies using hierarchical clustering, we developed Algorithm 1. We first discuss the algorithm in this section, and then in the following sections we offer some insight as to why such a simple algorithm is remarkably effective at creating taxonomies out of the noisy tag data generated by tens of thousands of users.

The algorithm starts with a single node tree whose only node is the "root" node representing the top of the tree (line 1). Then, it adds each tag in the tagging system to the tree in decreasing order of how central the tag is to the similarity graph described in the previous section (lines 2–4).[5] It decides where to put each candidate tag by computing its similarity to every node currently present in the tree, keeping track of the most similar node (lines 5–10). The candidate tag is then either added as a child of the most similar node if its similarity to that node is greater than some threshold, or it is added to the root node if there does not currently exist a good parent for that node (lines 11–16).

We have informally analyzed the output of our algorithm on several datasets, including the Delicious (output shown in Figure 1) and CiteULike tagging systems, discussed below. While there are occasional relationships in resultant trees which do not make sense, overall the algorithm seems to produce relatively consistent topical clusters of tags. Demonstrating success in something as large and qualitative as building a large hierarchy of thousands of nodes is difficult within the bounds of this note, though we will make available example hierarchies on our web site and are working on a user study to demonstrate the usefulness of our hierarchies both for navigation and narrowing/broadening browsing tasks.

Our algorithm is both fast and extensible. It is fast because using recent fast approximations of graph centrality (see [8] and [2]) it is possible to approximate certain types of centrality in a graph very quickly (especially in small world graphs, which seem to be produced commonly by tagging systems), and the rest of the algorithm's work consists of testing the similarity of each candidate tag to all potential parents in the tree at each phase of the algorithm. It is extensible because it is amenable to modifications to how candidate tags are attached to the growing hierarchy, for example: (a) adding the candidate tag to any and all tags which it is sufficiently similar to, (b) only adding the candidate tag if it is sufficiently similar to some subset of its ancestors, or (c) gradually reducing the required similarity threshold.

In the following sections, we investigate how our algorithm creates hierarchical taxonomies from tagging data.

[5] Various definitions of centrality exist from the social network analysis literature—we use closeness centrality.

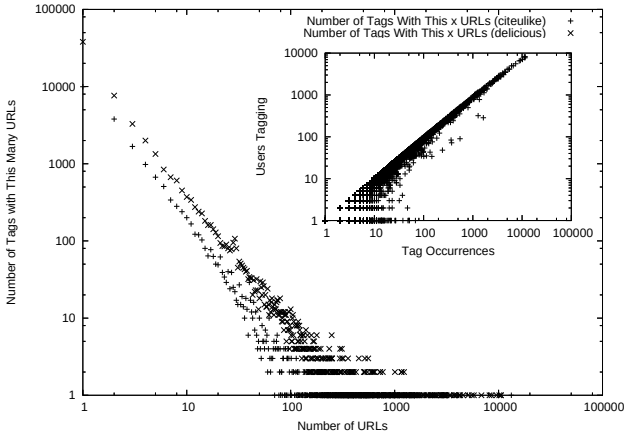


FIG. 2: The combination of a power law distribution of tags to objects in both Delicious and CiteULike (outset) and the strong correlation between the number of users using a particular tag and its popularity (inset) suggests that there is a strong sense of agreement on common tags among users.

In Section III, we describe the contents and differences between our tagging datasets, in Section III A, we describe the features which we believe significantly affect our algorithm, and in Section IV, we give a model which provides an explanation for why our algorithm works.

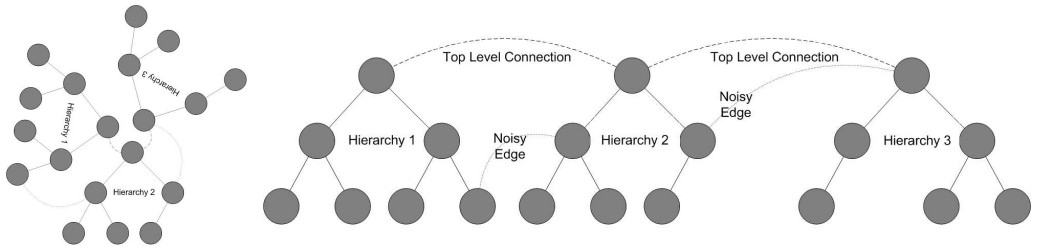
III. DATASETS

The data for our investigation consisted of two separate datasets from two substantially different tagging systems. Delicious [18] is a social bookmarking site with a community consisting primarily of technology and web developers. Our data from Delicious is a set of 19,594 distinct objects and their annotations gathered by saving all annotations associated with objects which were tagged during a chosen three day period. This dataset has 251,624 annotations by 84,349 distinct users with 60,219 distinct tags. In contrast, CiteULike [4] is a tagging system designed for scholars saving and sharing academic papers. Our CiteULike data is a set of 451,819 distinct objects and their annotations gathered by taking a random sample of all objects in the system. This dataset has 157,401 annotations by 5,732 distinct users and 41,523 distinct tags.

A. Data Features

We have determined several features of the tagging data which impact the effectiveness of our algorithm.

A prerequisite for generation of a tag hierarchy is that the data contains *natural hierarchical relations*. This seems to be a general feature of tagging data both because users appear to tag from their own personal mental taxonomies (leading to multiple levels on a per-user



(a) Similarity Graph

(b) Latent Hierarchical Taxonomy

FIG. 3: Our model suggests that there is a latent hierarchical taxonomy underlying the similarity graph in a collaborative tagging dataset, and that top level connections between general nodes in the hierarchy lead to centrality being a good generality metric.

basis) and because different users have different context-specific basic levels (the level of detail at which a user views an object).[6] An empirical study by Kome [13] shows that a large proportion of tags in Delicious participate in hierarchical relationships as defined in the appropriate ANSI/NISO and ALCTL taxonomy standards. A related concept to hierarchy is the *distribution of specificity* in the graph, or the level of detail of tags in the system.[7]

Agreement between users on a proper view of the world is critical for creating a useful shared context. In the tagging case, this translates into agreement on which tags are appropriate for a given subject. Figure 2 shows two graphs generated from our datasets suggesting that users do in fact agree to use the same tags to a large degree. The main graph shows power law distributions in the use of tags, suggesting that once one version of a tag (say "food" instead of "cuisine") becomes very popular, it is used even more. The inset shows a plot of the number of occurrences of a tag versus the number of users who use that tag. The distribution shows that in most cases, users only use a tag a few times, and that very popular tags are the product of thousands of users agreeing to use the same tag to label the same concept.

Density describes the frequency with which users annotate objects,

$$\frac{\text{annotated objects}}{\text{objects}}$$

and *overlap* describes the frequency with which users are annotating the same objects as one another,

$$\frac{\text{shared annotated objects}}{\text{annotated objects}}.$$

With respect to most of these features, the CiteULike data turns out to be more difficult to turn into a hierarchy: it has low density (some users do not tag objects at all), low overlap between users (academics working in different fields), and a specificity distribution which is very highly geared towards detailed tags. By contrast, Delicious is high density, high overlap, and has a much more even specificity distribution. While different algorithms might be better tuned for different specificity distributions, low density and overlap make it generally more difficult to create a hierarchy from tagging data (these factors mean that our algorithm needs a larger sample dataset from CiteULike than from Delicious to be effective).

IV. LATENT HIERARCHY MODEL

The features above suggest some aspects of tagging data to be considered in constructing new hierarchy building algorithms for tag data. However, the features by themselves do not suggest precisely why the algorithm presented works or what future steps might improve performance. We now sketch the rough model which has guided us on these points so far.

Our model of the tagging data, illustrated in Figure 3, concerns how similarities due to hierarchies in the data map to similarities in the similarity graph described in Section I B. Figure 3(b) shows the underlying hypothesized set of hierarchies which give rise to the similarity graph shown in Figure 3(a) that we actually analyze. In the model, we make three assumptions: (i) we hypothesize that the edges representing a given hierarchy also exist in the similarity graph (*hierarchy representation assumption*), (ii) that there are some noisy connections between tags that have no obvious relation to

[6] An example of differing basic levels is the case when a birder sees a "robin" when a normal person only sees a "bird." For more information see the discussion in [11] and the original [19].

[7] This can be understood as the number of objects from the differing levels of the hierarchy we are trying to generate—high specificity might indicate that there are many very detailed tags, like "chinesefood," but not very many broad tags, like "food."

one another in the underlying hierarchy (*noise assumption*), and (iii) that noisy connections between unrelated tags are more common higher up in a given hierarchy (*general-general assumption*). Without the first assumption, we have no way of detecting hierarchies using similarity (which is one of our few tools for analyzing tag relations). Any algorithm that ignores the second feature of the data is bound to give poor results because of the large amount of noise in both datasets. The third assumption seems to hold true in practice, though it may be be less universal than the other two. It is based on the intuition that in most cases, higher level tags are more likely to co-occur with one another by chance—"cat" co-occurring with "dog" is much more likely than "siamese" and "poodle."

The result of this model is that when the latent hierarchy of Figure 3(b) is translated into the similarity graph which we analyze, centrality becomes an effective measure of the generality of any single tag in the graph, so long as the hierarchies are reasonably well connected at the top by the general-general assumption. In doing so, we have shown that the social network notion of graph centrality seems to be as valid a way of determining importance in collaborative tagging systems as it is in social networks.

[1] G. Beenen, K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. E. Kraut. Using social psychology to motivate contributions to online communities. In *Proc. of CSCW '04*, 2004.

[2] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[3] J. J. Cadiz, A. Gupta, and J. Grudin. Using web annotations for asynchronous collaboration around documents. In *Proc. of CSCW '00*, 2000.

[4] R. Cameron. Citeulike. http://www.citeulike.org/, Mar. 2006.

[5] H. Chen. Collaborative systems: Solving the vocabulary problem. *IEEE Computer, Special Issue on Computer Supported Cooperative Work (CSCW)*, 27(5):58–66, 1994.

[6] C. Y. Chung, R. Lieu, J. Liu, A. Luk, J. Mao, and P. Raghavan. Thematic mapping - from unstructured documents to taxonomies. In *Proc. of CIKM '02*, 2002.

[7] K. C. Desouza. Facilitating tacit knowledge exchange. *Comm. ACM*, 46(6):85–88, 2003.

[8] D. Eppstein and J. Y. Wang. Fast approximation of centrality. In *Proc. 12th Symp. Discrete Algorithms*, pages 228–229. ACM and SIAM, 2001.

[9] G. W. Furnas. Experience with an adaptive indexing scheme. In *Proc. of CHI '85*, 1985.

[10] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987.

[11] S. Golder and B. A. Huberman. The Structure of Collaborative Tagging Systems, 2005.

[12] J. Hinrichs, V. Pipek, and V. Wulf. Context grabbing: Assigning metadata in large document collections. In *Proc. of ECSCW '05*, Paris, France, 2005. Springer.

[13] S. H. Kome. Hierarchical subject relationships in folksonomies. Master's thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, Nov. 2005.

[14] B. Lund, T. Hammond, M. Flack, and T. Hannay. Social bookmarking tools: A case study - connotea. *D-Lib Magazine*, 11(4), 2005.

[15] C. Marlow, M. Naaman, d. boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. 2005.

[16] C. C. Marshall and B. J. B. Brush. Exploring the relationship between personal and public annotations. In *Proc. of JCDL '04*, 2004.

[17] R. L. Riolo, M. D. Cohen, and R. Axelrod. Evolution of cooperation without reciprocity. *Nature*, 414(6862):441–443, 2001.

[18] J. Schachter. del.icio.us. http://del.icio.us/, Mar. 2006.

[19] J. Tanaka and M. Taylor. Object categories and expertise: is the basic level in the eye of the beholder? *Cogn. Psychol.*, 23:457–482, 1991.

[20] C. Weng and J. H. Gennari. Asynchronous collaborative writing through annotations. In *Proc. of CSCW '04*, 2004.

[21] H. Wu, M. D. Gordon, and K. DeMaagd. Document co-organization in an online knowledge community. In *Proc. of CHI '04*, pages 1211–1214, 2004.

[22] H. Wu, M. D. Gordon, K. DeMaagd, and N. Bos. Link analysis for collaborative knowledge building. In *Proc. of HYPERTEXT '03*, pages 216–217, 2003.

[23] Yahoo gobbles up delicious. Reuters, Dec. 2005.