Inducing Ontology from Flickr Tags

Patrick Schmitz
University of California, Berkeley
and
Yahoo! Research, Berkeley
pschmitz@sims.berkeley.edu

ABSTRACT
In this paper, we describe some promising initial results in inducing ontology from the Flickr tag vocabulary, using a subsumption-based model. We describe the utility of faceted ontology as a supplement to a tagging system and present our model and results. We propose a revised, probabilistic model using seed ontologies to induce faceted ontology, and describe how the model can integrate into the logistics of tagging communities.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Retrieval models, Search process*,
H.3.1 [**Content Analysis and Indexing**]: *Abstracting methods, Indexing methods*,
H.3.5 [**Online Information Services**]: *Data Sharing*,
G.3 [**Probability and Statistics**]: Correlation.

Keywords
Tagging, Subsumption, Ontology, Facets.

1. INTRODUCTION
In the last few years, we have seen rapid growth tagging applications, both in the number of applications leveraging a tagging model, as well as in the number of users participating in tagging communities. This growth currently outpaces our understanding of how to make the resulting annotations efficient and productive for a range of uses and users.

Tagging systems are often placed in opposition to taxonomic models, and two issues are commonly cited:

1. Annotation user interfaces based upon a closed, hierarchical vocabulary are awkward and inflexible;
2. A strict tree of concepts does not reflect usage and intent.

The first criticism is valid, but can readily be addressed with dynamic ontologies and better UI mechanisms. Much of the second criticism is not so much an issue with taxonomy (ontology) per se, but rather with problematic models that force users to place concepts into a single hierarchy. Much of this issue can be addressed using faceted ontologies that separate the different aspects of annotation intent.. Some of the common facets used in media annotation include location, associated activity or event, various depiction facets (people, flora, fauna, objects et al.) and especially in the sharing context, emotional response.

Tags provide a simple and direct mechanism to create annotations that reflect a variety of facets, and also provide a direct means of embarking upon a search. However purely tag-based search tends to have low recall performance (this can be partially mitigated with a UI that encourages aligned vocabulary). Moreover, when an initial search returns a large number of results, tags do not support efficient or intuitive query refinement models. At best, users can currently refine a search using clusters of (statistically) related concepts. Although sometimes useful, clustering performance is very difficult to evaluate.

Sanderson and Croft [7] describe the distinction between polythetic clusters (in which cluster members share some proportion of a set of characteristics) and monothetic clusters in which all members share a common feature. They argue (and we concur) that users can more easily understand monothetic clusters. In addition, where the polythetic clusters are difficult to label, easily labeled monothetic clusters lend themselves to various common interface paradigms such as guided navigation, sparse hierarchies for query refinement, et al.

Dumais et al. [2] have explored faceted search mechanisms, and Hearst [4] and Yee, et al. [8] demonstrate the utility of faceted search interfaces for image search. Although some in the tagging community resist taxonomy, even del.icio.us (http://del.icio.us) recently added "bundles" - while purely organizational (bundles do not support ontological semantics), the feature acknowledges the organization problems of scaling a tagging model.

We believe that users should not have to choose between pure tag-based models and pure taxonomic models with closed vocabularies. We are exploring a model that leverages statistical natural language processing techniques together with domain knowledge to induce ontology that can be leveraged on the back-end. Our objective is a system that preserves the flexibility of the tagging interface for annotation while also benefiting from the power and utility of a faceted ontology in the search and browse interface.

We present early results of a subsumption based model on the Flickr (http://www.flickr.com) tag set, demonstrating the potential for such a technique to induce ontology suitable for a search and browse user interface. The rest of the paper describes the approach, the data set and evaluation, and a proposal for a refined model and how this would fit into the logistics of a tagging community like Flickr.

2. RELATED WORK
Sanderson and Croft describe a simple statistical model for subsumption in which X subsumes Y if:

```
P(x|y >= 0.8) and P(y|x < 1)
```


WWW 2006, May 22–26, 2006, Edinburgh, UK.

They apply this co-occurrence model to concept terms extracted from documents returned for a directed query (using query results helps to constrain the domain of terms). Clough et al. [1] adapt the same technique to curated (i.e., professionally annotated) photographs from an historical collection. The resulting taxonomies are fairly noisy (i.e., many of the proposed subsumption pairs are incorrect), especially given that the domain vocabularies are focused by the original queries. We include their results in Table 1 below as a baseline. Despite the low yield, these models generate taxonomy that reflects the actual usage, and so are well suited to tagging applications.

Many others have experimented with inducing ontology using statistical NLP techniques, including Hearst [3], Yee, et al. [8] and Mani, et al [5]. Some of these [3], [5] depend at least in part upon grammatical speech, and so can only be applied in natural language contexts. Others [8] attempt to match concepts to existing ontologies such as WordNet; these models may be inherently less noisy, but since WordNet is based upon standard English vocabulary it may be difficult to adapt such models to the dynamic and sometimes idiosyncratic vocabulary that emerges in tagging applications (e.g., for event names).

3. EXPLORATORY APPROACH

3.1 Subsumption step

We adapted the model of [7] to the Flickr tag set, adjusting the statistical thresholds to reflect the ad hoc usage, and adding filters to control for highly idiosyncratic vocabulary. Thus X potentially subsumes Y if:

```
P(x|y >= t) and P(y|x < t),
 Dx >= Dmin, Dy >= Dmin,
 Ux >= Umin, Uy >= Umin
```

Where:

- $\mathbf{t}$ is the co-occurrence threshold,
- $\mathbf{D_x}$ is the # of documents in which term $\mathbf{x}$ occurs, and must be greater than a minimum value $\mathbf{D_{min}}$, and
- $\mathbf{U_x}$ is the # of users that use $\mathbf{x}$ in at least one image annotation, and must be greater than a minimum value $\mathbf{U_{min}}$.

We filter the input documents (i.e., the photos), requiring a minimum of 2 tag terms so that co-occurrence was defined.

We conducted a series of experiments, varying the parameters $\mathbf{t}$, $\mathbf{D_{min}}$, and $\mathbf{U_{min}}$. We were looking for a balance that minimized the error rate and maximized the number of proposed subsumption pairs.

Using stricter values for the co-occurrence threshold (approaching 0.9) reduces the error rate somewhat, but drastically reduces the number of proposed pairs. Useful values were between 0.7 and 0.8, somewhat below the comparable value determined empirically by Sanderson and Croft [7].

The model is more sensitive to changes in $\mathbf{U_{min}}$, than $\mathbf{D_{min}}$. Setting $\mathbf{U_{min}}$ to anything below 5 produced many highly idiosyncratic terms in noisy subsumption pairs; a useful range was 5 to 20. $\mathbf{D_{min}}$ values varied from 5 to 40, and proved to be useful as a means of fine tuning. Both values were increased slowly as the number of documents was increased. With input sets below 1 million photos the vocabulary was less stable and so the model was more sensitive to the parameters.

3.2 Tree pruning and reinforcement

Once the co-occurrence statistics are calculated, candidate term pairs are selected using the specified constraints. We then build a graph of possible parent-child relationships, and filter out the co-occurrence of nodes with ascendants that are logically above their parent.

I.e., for a given term $\mathbf{x}$, and two potential parent terms $\mathbf{p_{xi}}$ and $\mathbf{p_{xj}}$ if $\mathbf{p_{xi}}$ is also a potential parent term of $\mathbf{p_{xj}}$ then we remove $\mathbf{p_{xi}}$ from the list of potential parent terms for term $\mathbf{x}$. At the same time, the co-occurrence of terms $\mathbf{x}$, $\mathbf{p_{xi}}$ and $\mathbf{p_{xj}}$ in the given relationships indicates both that the $\mathbf{x->p_{xj}}$ relationship is more likely than simple co-occurrence might indicate, and similarly that the $\mathbf{p_{xi}->p_{xj}}$ relationship should be reinforced. We increment the weights of each accordingly. Finally, we consider each leaf in the tree and choose the best path up to a root, given the (reinforced) co-occurrence weights for potential parents of each node, and coalesce the paths into trees[1].

With sufficiently large document sets, many of the resulting trees are fairly broad - e.g., cities with points of interest (see the Evaluation section below for a discussion of how we define subsumption relationships). We noted a disproportionate number of erroneous paths in singleton and doubleton subtrees, as compared to the larger subtrees, and so filter these out altogether. This is justified given that the total number of candidate trees was very large for these runs (from 2000 to 6000+ candidate pairs meeting the basic subsumption and filtering criteria), and since the ultimate goal is to provide enough structure to assist sensemaking and guide navigation through the collection. A secondary goal is to improve search by inferring parent terms for images with child terms, and in this sense some recall is certainly sacrificed in filtering out the singleton and doubleton trees. We believe that users of the subsumption trees will be more sensitive to precision than recall, but this aspect of the model must be evaluated in large scale user studies.

4. DATA SET AND ANALYSIS

We used a snapshot of the Flickr metadata database as of July 2005. At this point, there were some 25 million total images uploaded, and roughly 65 million total annotations. Roughly 5 million of these images were marked as "not public" and so were excluded from the experimental set. The tables were modified to anonymize all user data (including photo ids) and all images with fewer than 2 terms were filtered. This resulted in a data set of roughly 9 million images. The associated vocabulary has well over 200K terms and over 8 million total pairs (a precise number is not available as we filter some as we go to reduce memory overhead).

Among the Flickr annotations, the vocabulary is inconsistent with respect to both spelling and term boundaries (e.g., "San Francisco" often shows up as two terms "san" and "francisco", due to a somewhat non-intuitive tag entry interface). In addition, there are many idiosyncratic annotation terms. These latter vary from personal events described as a phrase in a tag ("johnandmaryswedding" – possibly indicating some confusion

[1] This has the potential to conflate multiple meanings of a given term, and in some cases does cause sensible lower paths to be "grafted" onto inappropriate higher paths. However, the number of errors introduced in this manner is quite small.

Table 1. Performance Summary

Model	# rel'ns, avg.	correct	related /aspect	same*	error /other
Sanderson and Croft	?	23%	49%	8%	19%
Clough et al.	105	15%	10%	0.2%	43%
Our model	1200+	51%	21%	5%	23%

about the tagging model, or some conflation of tagging with description) to slang, abbreviations and Flickr-cultural curiosities ("thrash", "deleteme997", "c").

4.1 Subsumption - evaluation

The resulting trees were evaluated manually. Each proposed subsumption pair was marked as correct, inverted, related, synonymous (including language variants on common terms like "flower"/"blume"/"fleur"/"bloem" etc.), or noise (wholly erroneous).

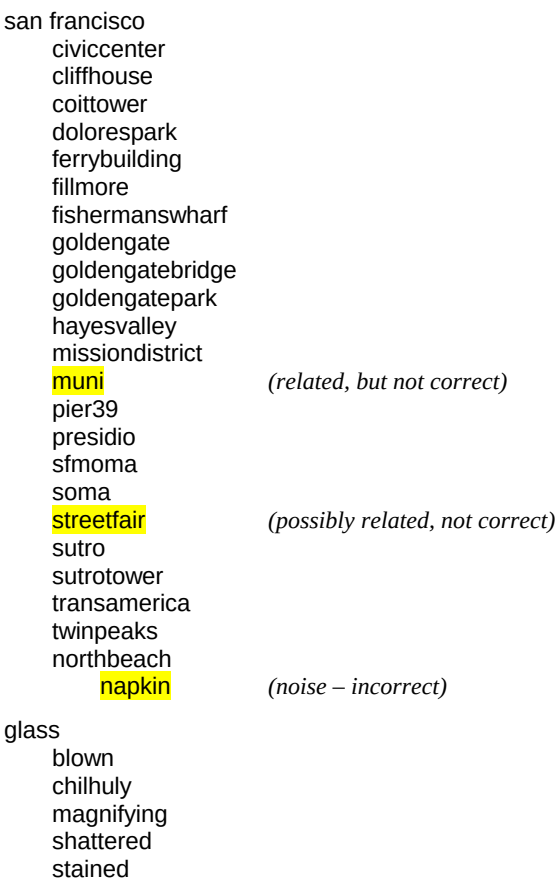
Figure 1 shows several examples of the types of trees generated. Many of the child concepts of “San Francisco” are neighborhoods or points of interest; several are (possibly) related and there is one example of the noise that results from a statistical model. In the second example, each of the child nodes is a hyponym of “glass”; although perhaps not what an art historian would create as a model of the domain, it is ‘correct’ in representing the usage within the Flickr community.

Based upon our experience and that of others (e.g., Naaman et al. [6]), we hypothesized that images will be annotated and most easily retrieved when emphasizing several key facets: place, activity and depictions. The Flickr community also seems to emphasize another facet that might best be described as *emotion* or *response*. In our results a large proportion of the shared vocabulary is tied to placenames, although we expect that model refinements will produce more of a balance with other facets.

For location, we consider a combination of geographic placenames as well as points of interest that demarcate place more than activity. Thus we consider "San Francisco" a reasonable parent of "Golden Gate Bridge". In the sense of a pure type-of relationship this would not hold, however for the utility of locating an image, it is entirely reasonable. By the same token "San Francisco" may be related to, but is not a parent of "muni" or "streetfair". For generic terms like "lake" and "park", we considered instances of lakes or parks to be reasonable children. In depictions, more typical type-of relationships were used: "dog" subsumes specific breeds, "food" subsumes "kimchee" and "creamcheese" where "restaurant" is only related. In a large photo sharing environment such as Flickr, personal relations are less useful for query, and so we regard nearly all personal names as noise in any pair context.

Table 1 compares the results for related subsumption models to our results. Sanderson and Croft [7] report high "aspect of" numbers, and ascribe this at least in part due to the vocabulary-limiting queries. Clough et al. [1] is a similar application to ours and so provides a useful baseline. We believe their model would perform better if applied to the entire vocabulary of their dataset rather than a focused query. Both [7] and [1] seem to contain an inconsistency in the statistical model (the second term should be expressed as `P(y|x < 0.8)` and not `P(y|x < 1)`), although this may be a typographical error in the papers.

Figure 1. Example output



5. PROPOSED MODEL

The initial results are promising enough to prompt further work. Our model produces subtrees that generally reflect distinct facets, but cannot categorize concepts into facets. We have planned a series of changes to the model to address this.

5.1 Move to a pure probabilistic model

We are currently working to express subsumption, pruning and tree construction, and facet categorization together in a unified probabilistic model, somewhat along the lines of Mani et al. [5]. A probabilistic model should be more robust, and incorporate concepts like “the number of authors using a tag” as a feature scale rather than a simple threshold as in the current model.

5.2 Add deduplication/misspelling support

We also would like to add better support for deduplication and misspellings; we believe the current Flickr user interface produces more of these than models that support tag suggestion (e.g., del.icio.us). By representing the resulting ontology as a graph of concepts that have various labels, we can associate variant

spellings in a probabilistic manner. The most common spelling is the natural label.

5.3 Explore morphological tools

We are also exploring morphological analyses, although we are concerned about the potential to conflate facets. Early analysis of the data indicates that certain morphological techniques (e.g., depluralization and verb-gerund stemming) may be appropriate to some facets and not to others.

5.4 Seed with faceted ontologies

A significant problem with subsumption is that in common usage, people tend to name generic concepts - neither too general nor too specific. In particular, users rarely specific generic concepts such as "country" or "continent" for location, and "mammal" or "plant" for depiction. In our results, for example, certain country names were rarely specified and so were placed under cities. However, these upper level ontological concepts are freely available in the form of gazetteers and common taxonomies. We plan to seed our new model with these domain-specific upper model ontologies (DUMO's). This will address the inherent weakness in subsumption, but it serves another purpose as well. By specifying the top level structure of the ontology, we can enforce the facet model that makes the most sense for users; since it is an input into the model, we can easily test variants on this with the user base.

5.5 Support community moderation

While we expect the refined model to reduce the noise (errors) in our results, we believe that the model may best be deployed not as a fully automated process, but rather as a productivity tool. Many tagging applications have an established model for community, including enthusiast moderators for popular sub-domains.

If the statistical model can suggest ontology to a set of moderators, they need only approve or reject the proposed relations. Once a baseline is established, it requires little effort from the moderators to keep the ontology fresh, reflecting current usage. The statistical model reflects community usage, with moderators acting as a check and balance.

6. CONCLUSIONS

We have described a subsumption based model for inducing ontology from tag usage that produces promising initial results. Refining this model we hope to improve upon the accuracy, and also to induce faceted ontology. The results will support more effective search and browser interfaces, and can be reasonably integrated into existing community models by leveraging enthusiasts as moderators.

7. ACKNOWLEDGMENTS

Our thanks Yahoo Inc. for the use of the Flickr data, and to Prof Dan Klein, U.C. Berkeley, for his advice on the model.

8. REFERENCES

[1] Clough, P., Joho, H. and Sanderson, M. (2005) “Automatically Organising Images using Concept Hierarchies”. In: Proceedings of Multimedia Information Retrieval 2005

[2] Dumais, S, et al. "Stuff i've seen: A system for personal information retrieval and re-use". In SIGIR, 2003.

[3] Hearst, M., (1992) Automatic Acquisition of Hyponyms from Large Text Corpora, in “Proc. of COLING 92”, Nantes.

[4] Hearst, M. (1999). "User Interfaces and Visualization". In: Baeza-Yates, R. & Ribeiro-Neto, B. (eds.), Modern Information Retrieval, pp. 257-323. New York: ACM Press.

[5] Mani, I., Samuel, K., Concepcion, K., and Vogel, D. "Automatically Inducing Ontologies from Corpora". Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology, COLING'2004, Geneva.

[6] Naaman, M, et al. "Context Data in Geo-Referenced Digital Photo Collections". In proceedings, Twelfth ACM International Conference on Multimedia (ACM MM 2004), October 2004.

[7] Sanderson, M. and Croft, B. (1999) “Deriving concept hierarchies from text” In: Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval, pp. 206-213.

[8] Yee, K-P., Swearingen, K., and Hearst, M. (2003) “Faceted metadata for image search and browsing”. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 401-408.