Constructing Folksonomies from User-Specified Relations on Flickr

Anon Plangprasopchok
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
plangpra@isi.edu

Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
lerman@isi.edu

ABSTRACT

Automatic folksonomy construction from tags has attracted much attention recently. However, inferring hierarchical relations between concepts from tags has a drawback in that it is difficult to distinguish between more popular and more general concepts. Instead of tags we propose to use user-specified relations for learning folksonomy. We explore two statistical frameworks for aggregating many shallow individual hierarchies, expressed through the collection/set relations on the social photosharing site Flickr, into a common deeper folksonomy that reflects how a community organizes knowledge. Our approach addresses a number of challenges that arise while aggregating information from diverse users, namely noisy vocabulary, and variations in the granularity level of the concepts expressed. Our second contribution is a method for automatically evaluating learned folksonomy by comparing it to a reference taxonomy, e.g., the Web directory created by the Open Directory Project. Our empirical results suggest that user-specified relations are a good source of evidence for learning folksonomies.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data mining*; I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning—*Knowledge Acquisition*

General Terms

Algorithms, Experimentation, Human Factors, Measurement

Keywords

Folksonomies, Taxonomies, Collective Knowledge, Social Information Processing, Data Mining

1. INTRODUCTION

The Social Web is changing the way people create and use information. Unlike traditional Web sites, Flickr, Digg, YouTube, among many others, allow users to create, organize and distribute many different types of content, including images, news stories, videos, and maps. In the course of creating and using content, users often annotate it with metadata in forms of discussions, ratings, descriptive labels known as *tags*, and links between content, metadata and users. The collective knowledge expressed though user-generated, user-annotated data has the potential to transform many fields, including information discovery [9], management of the commons [20], and even the practice of science [18]. In order to leverage the collective knowledge, we need tools to efficiently aggregate data from large numbers of users with highly idiosyncratic vocabularies, varying degrees of expertise, and who are governed by different, sometimes conflicting incentives [12, 5].

A taxonomy is a hierarchical classification system used to organize our knowledge of the world. The Linnean classification system, one of the best known taxonomies, is used to categorize all living organisms. Other examples of taxonomies (not necessarily strictly hierarchical) include library classification schemes, e.g., the Dewey Decimal system, and Web directories that categorize Web pages, e.g., the Yahoo directory. The explosion of social metadata has led to several efforts [13, 17] to learn a common informal taxonomy — a so-called *folksonomy* — from the tags used by large numbers of users to annotate content for their personal use. Unlike a formal taxonomy created by a small group of *experts* using a *controlled vocabulary*, a folksonomy emerges bottom-up from the bits of knowledge about the world expressed by many users using *uncontrolled* personal vocabularies. The advantages of an automatically learned folksonomies are that they are relatively inexpensive to produce, dynamic, evolving in time as community's needs and vocabulary change, and can be used to improve information search and discovery (e.g., [15]).

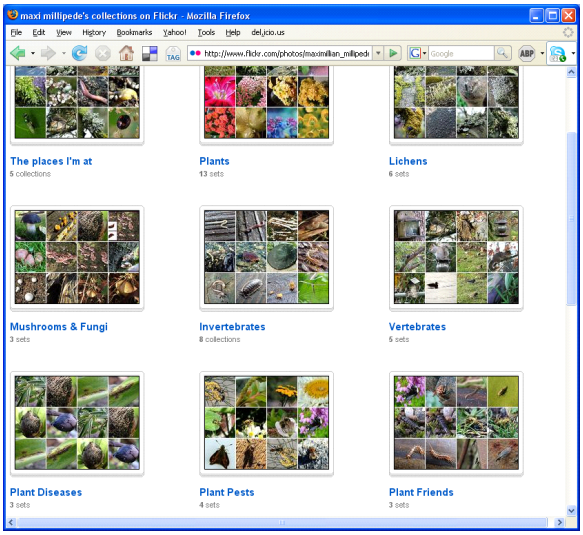
Current approaches to automatic folksonomy construction combine tags created by distinct individuals using statistics of their co-occurrence [17, 7, 22]. However, we believe that attempts to learn hierarchical "broader/narrower" relations between concepts using tag frequency alone will not be able to properly distinguish between popular and general concepts. For instance, there are ten times as many images on the photosharing site Flickr tagged with "car" than with "automobile", a concept that subsumes "car." Instead of tags, we use a novel source of evidence, *user-specified relations*, to learn a common folksonomy. Recognizing that tags may not be sufficiently expressive to annotate a variety of content, some social web sites have began to allow users to organize their metadata and content hierarchically. The social bookmarking site Del.icio.us, for example, allows users to manually group related tags into *bundles*, while Flickr allows users to group related photos into *sets* (i.e., photo albums), and related sets into *collections* (and related collections in other collections). Although these sites

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

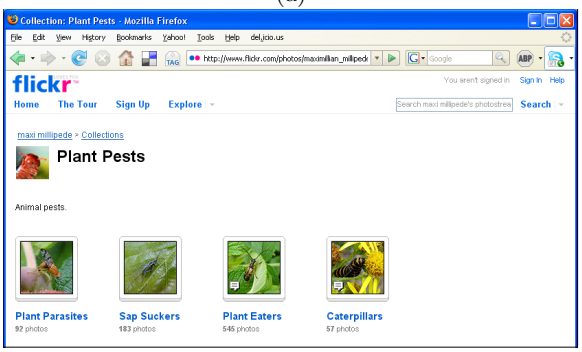
do not impose any constraints on the hierarchies, we find that users employ them to specify *relations* between concepts, specifically "broader/narrower" relations. We claim that user-specified relations are a good source of evidence for learning folksonomies.

In this paper, we present a statistical framework for aggregating many shallow individual hierarchies, expressed through the *collection/set* relations on Flickr, into a common folksonomy that reflects how a community organizes knowledge. Our approach addresses a number of challenges that arise while aggregating information from diverse users. Noise is an issue in this data, since users' vocabulary can be highly idiosyncratic. Another challenge is individual differences in the level of expertise and granularity: one user may organize photos first by country and then by city, while another organizes them by country, then subregion and then city. Aggregating data from these users may potentially generate multiple paths from one concept to another. Determining which path to be retained is a non-trivial problem. Yet another challenge is variation in the classification order used. Suppose user A organizes her photos by activity, e.g., creating a collection she calls travel first, and as part of this collection, a set called china for photos of her travel in China. Meanwhile, user B organizes her photos by location, creating a collection china, with constituent sets travel, people, food, etc. Both schemes are correct, and a folksonomy learning method must be able to deal with them.

The contributions of this paper are three-fold. First, we describe user-specified relations, and how they are used on Flickr (Section 2). We argue that this metadata constitutes a novel source of evidence for learning folksonomies. Second, we present simple, yet intuitive, statistical frameworks for selecting meaningful relations and joining them in a folksonomy (Section 3). We present empirical results of folksonomy learned from Flickr data in Section 4. In particular, we present a method for automatically evaluating the learned folksonomy by comparing it to the web directory maintained by the Open Directory Project (ODP).



(a)



(b)

Figure 1: Personal hierarchy specified by a Flickr user. (a) Some of the collections created by the user and (b) sets associated with a specific collection.

2. USER-SPECIFIED RELATIONS

In addition to "flat" keywords or tags, some social Web sites have recently began to provide a feature enabling users to organize content hierarchically. While the sites themselves do not impose any constraints on the vocabulary or the semantics of the relations used, in practice users employ them to represent both subclass relationships (`dog` is a kind of `mammal`) and part-of relationship (`my kids` is a part of `family`). Users appear to express both types of relations through the hierarchy, in effect using the hierarchy to specify broader/narrower relations. Even without strict semantics being attached to these relations, we believe that user-specified relations represent a novel source of evidence for learning folksonomies that is superior to using tags alone. We describe how the social photosharing site *Flickr*[1] implements this feature.

Flickr allows users to group their photos in album-like folders, called *sets*. Users can also group sets into "super" albums, called *collections*.[2] Both sets and collections are named by the owner of the image. A photo can be part of multiple sets. It can also be submitted to any of the thousands of special-interest groups Flickr users have created to share photos on a given topic.

Flickr does not enforce any specific rules about how to organize photos in sets and collections or how to name them. While some users create multi-level hierarchies containing collections of collections, etc., the vast majority of users who use collections create shallow hierarchies, consisting of collections at the top level and their constituent sets. We found that most users group "similar" or "related" photos into the same set, and group related sets into the same collection. Figure 1(a) shows collections created by an avid naturalist on Flickr. These collections reflect the subjects she likes to photograph: *Plants*, *Mushrooms & Fungi*, *Invertebrates*, *Plant Pests*, etc. Figure 1(b) shows the constituent sets of the *Plant Pests* collection: *Plant Parasites*, *Sap Suckers*, *Plant Eaters*, and *Caterpillars*. The name of the set generally subsumes all the photos within it (e.g., the *Caterpillars* set contains photos of caterpillars), while the collection name is usually broad enough to cover all the sets within it (`caterpillars` and `sap suckers` are types of `plant pests`). In general, users seem to employ collection-set hierarchy to express broader-narrower relations.

[1] `http://www.flick.com`

[2] The collection feature is limited to paid "pro" users. Pro users can also create unlimited number of photo sets, while free membership limits a user to three sets.

3. AGGREGATING RELATIONS FROM DIFFERENT USERS

We define C^i and S^{ij} as a collection i and a set j of the ith collection respectively.[3] A collection or set name contains a series of terms: $< t_1, ..., t_k >^{C^i}$ is a name of C^i and $< t_1, ..., t_l >^{S^{ij}}$ is a name of S^{ij}.

As discussed above, we assume that relations that a user specifies through collections and sets are broader-narrower type relations. We denote that C^i is broader than S^{ij} as $C^i \rightarrow S^{ij}$. These relations are also applicable to their constituent terms (relation delegation). In particular, if a user specifies the set S^{ij} under the collection C^i — the former is narrower than the latter, and all the terms in S^{ij} are also narrower than those of C^i. We also assume that each of those terms represents a concept in a conceptual hierarchy, and that the same terms used by the same or different users represent the same concept.[4]

There are three main steps involved in learning folksonomies from user-specified relations: (1) data preprocessing step that extracts and normalizes terms; (2) relation weighting and pruning; (3) concept integration that links shallow hierarchies into a common deeper hierarchy. We explore two statistical frameworks for picking meaningful relations and integrating many shallow hierarchies from different users into a common deeper hierarchy. The first framework identifies relations that have the highest agreement. These relations are then linked into a deeper folksonomy. This method could potentially lead to formation of multiple paths between concepts. We cast this multiple path problem as maximum bottleneck path, which provides a method to select a path corresponding to the highest agreement. The second approach identifies the most informative, or significant, relations, i.e., those that are highly unlikely to be observed purely by chance in the data. Significant relations are then linked into a deeper folksonomy. If there exist multiple paths between nodes, only the longest one is retained. We also describe a subsumption-based approach which infers broader-narrower relations from co-occurrence probabilities of the terms in collection and set names. This method was previously used to learn folksonomies from tags [17].

We will briefly describe data preprocessing first since this step is shared across different frameworks. Steps 2 and 3 are described separately under each framework.

3.1 Data Preprocessing: Term Extraction and Normalization

First, we extract terms representing concepts from collection and set names. We found that users often combine two or more concepts within a single name by using words and special characters to join different concepts, e.g., "Dragonflies/Damselflies", "Mushrooms & Fungi", "Moth at Night." These bridge words include prepositions, such as "at", "of", "in," and conjunctions, such as "and" and "or." The special characters include '&', '<', '>', ':', '/'. We start by tokenizing collection and set names on these words and characters. We do not tokenize on white spaces to avoid breaking up composite terms like "South Africa." We remove terms composed only of non-alpha numeric characters and frequently-used uninformative words, *e.g.*, "me" and "myself." We then lowercase all terms and use the Porter stemming algorithm to normalize the remaining terms. This step is necessary to mitigate noise due to individual variations in naming conventions and vocabulary usage.

Once terms are extracted and normalized, each unique term is treated as a concept, and concept relations are delegated from collection-set relations. Thus, if in our data set we have a collection named "Odonata" with a set named "Dragonflies/Damselflies", we create two relations: *odonata* → *dragonfli* and *odonata* → *damselfli*.

After normalizing data, we remove overly vague or overly specific concepts and relations. We discard relations that are used only by a single user. Extracted relations are also used to remove concepts that are too broad to be useful, *e.g.*, "all set", "all my set", "world travel." In particular, we use the ratio between a number of child and parent concepts to determine whether a concept is uninformative. A concept with a high ratio covers too many concepts, while having very few or no concept covering it. In this study, we discard top 100 highest ratio concepts.

3.2 Relation Weighting and Linking

Once relations are extracted, the next step is to aggregate and link them together into deeper hierarchies. Since each relation is extracted from different users' collection-set relations, our data set is very noisy due to idiosyncracies in users' categorization schemes, differences in opinions, vocabulary, level of expertise and so on. For example, there are 30 users who express *europ* → *itali*, and one user who expresses *itali* → *europ*. Moreover, relations from different users, when aggregated and linked, can result in multiple paths from one concept to another, e.g. relations *anim* → *bug*, *bug* → *moth*, and *anim* → *moth*, resulting in two different paths between *anim* and *moth*. However, since the longer path subsumes the shorter path, while providing an addition level of detail, it should be retained, with the shorter path dropped, to simplify the learned folksonomy. In this section, we describe approaches that address these issues.

We propose two statistical frameworks to weight shallow relations and then link them together into deeper folksonomies. We also briefly describe probabilistic subsumption approach, which was previously used for inducing shallow hierarchies from tags [17]. This approach will be used as a baseline.

3.2.1 Conflict Resolution Framework

The basic premise of this approach is that relation conflicts occur because of noise, when a minority of users specify relations opposite to those of the majority. For each relation, we simply consider how many users agree and disagree on it, i.e., how many users express forward and backward relations for a certain concept pair. Intuitively, concept a subsumes (or is broader than) concept b if a number of users who agree upon $a \rightarrow b$ is greater than the number who agree on $b \rightarrow a$, with some threshold:

let $d_{x \rightarrow y}$ be the number of users who define $x \rightarrow y$
and $d_{y \rightarrow x}$ be the number of users who define $y \rightarrow x$
We define x "subsumes" y over all users if:
$d_{x \rightarrow y} > 1$ and
$d_{y \rightarrow x} < d_{x \rightarrow y}$

Where conflicts exist, we use a majority opinion to find and retain meaningful relations, and discard conflicting relations expressed by a minority of users.

[3] A collection and its sets are specific to an individual user.
[4] Although polysemy and synonymy do exist on Flickr, we ignore them for reasons of simplicity in this paper.

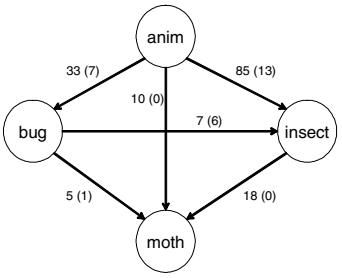


Figure 2: An illustrative diagram represents relations (arrows) between four concepts (circles): `anim`, `insect`, `bug`, and `moth`. The numbers represent the number of users who agree (disagree) on a particular relation, e.g., `anim` → `bug` (vs `bug` → `anim`).

Although conflict resolution helps filtering out "noisy" relations, it does not address the issue of multiple paths from one concept to another. This issue is partly caused by the varying levels of specificity used by different users, and also by users' categorization variation. As an example, some users define `anim` → `insect` and/or `insect` → `moth`, while others define `anim` → `moth` directly, as shown in Figure 2. As mentioned earlier, multiple paths may lead to aggregated relations being densely linked, making the learned folksonomy unnecessarily complex and hard to use. We need an approach to determine which paths should be kept and which discarded.

Since a path is composed of relations with different weights (numbers of users who express such relations), one way to score this path is to use the minimum weight among these relations. This minimum weight can be cast as Network Bottleneck in Network Optimization problems [2]. Basically, we view each concept as a node, a relation as an edge and a number of users who agree on a certain relation as a information-flow capacity, or the weight, of that edge. For a certain path from one concept to another, we determine flow bottleneck. The flow bottleneck is a minimum flow capacity among all relations (edges) in the path. This bottleneck score will be used to score the path. Intuitively, it measures the amount of users' agreement on a path. After scoring all possible paths, the path with the least disagreement will be chosen.

This process can be formally described as follows. Given source a and sink b concepts,

$$\max_i(P^i_{a\to b}) = \max_i(\min_j\{W(e_{ij})|e_{ij} \in E(P^i_{a\to b})\}),$$

where $P^i_{a\to b}$ is a path i from concept a to b, e_{ij} is relation j of the path i; $E(x)$ is a function returns all relations in the path x, and $W(y)$ returns the weight of the relation y. Considering the case in Figure 2, the bottleneck score for `anim`→ `insect` →`moth`is 18 (we subtract a number of conflicting relations); `anim`→`moth` is 10; `anim` → `bug` → `moth` is 4; `anim` → `bug` → `insect` → `moth` is 1. Consequently, `anim` → `insect` → `moth` is chosen.

3.2.2 Significance Test Framework

This approach finds meaningful relations in the data by checking whether they are statistically significant. Consider a particular relation from concept a to b. We use hypothesis testing approach to decide whether a relation $a \to b$ is significant, i.e., highly unlikely to arise purely by chance in our data. In this context, the null hypothesis is that observed relations were generated by chance, via the random, independent generation of the individual concepts. Hypothesis testing decides, at a given confidence level, whether the data supports rejecting the null hypothesis. Suppose n instances of a concept a were generated by a random source. The probability that a concept b (which occurs with an overall probability p in the data) was used as a child of a k times has a binomial distribution. We will reject the null hypothesis if k is larger than was expected if relations were generated by chance.

In order to determine if k is large enough for rejecting the null hypothesis, we first compute cumulative probability of the binomial distribution, i.e., the probability of observing at least k events. For a large n, the binomial distribution approaches a normal distribution $N(x, \mu, \sigma)$ with $\mu = np$ and $\sigma^2 = np(1-p)$. The cumulative probability in observing at least k events is:

$$p(x \geq k) = \int_{x=k}^{\infty} N(x, \mu, \sigma)dx\,. \quad (1)$$

We approximate the value of the integral in (1) using approximation formulas in [1].

The significance level of the test, α, is the probability that the null hypothesis is rejected even though it is true, and it is given by the cumulative probability above. Suppose we set $\alpha = 0.01$. This means that we expect to observe at least k events 1% of the time under the null hypothesis. If the number of users who expressed the relation $a \to b$ is greater, we reject the null hypothesis, i.e., decide that the relation is significant.

After discarding all uninformative relations using significance testing approach, we still need to select the best path out of several possible ones linking one concept to another. Since all retained relations are judged to be significant, we cannot rank paths using Network Bottleneck metric as in the *Conflict Resolution* framework. Instead, we simply select the longest path. In the example in Figure 2, suppose that all relations are significant. Then, the path *anim* → *bug* → *insect* → *moth* will be selected.

3.2.3 Term Subsumption Framework

As a baseline for this study, we apply the probabilistic subsumption approach to induce shallow relations. Basically, we create bags-of-terms from the terms used in collection and set names. Each bag represents a given set and is composed of terms from the names of the set and the collections to which the set belongs. Although subsumption approach was originally applied to learn a folksonomy from Flickr images annotated with descriptive tags [17], we believe that using terms in collection and set names will have the same effect. In particular, terms from collections will appear in the bags more frequently than those from the sets; therefore, the former will subsume the latter. A benefit of using the same data for the relation-based and subsumption-based approaches is that the folksonomies are learned from the same vocabulary, making direct comparison feasible.

Following Sanderson and Croft [16], term occurrences and co-occurrences are used to determine if one term subsumes another term. The term occurrence of a is computed from the number of all bags-of-terms in which a appears; and

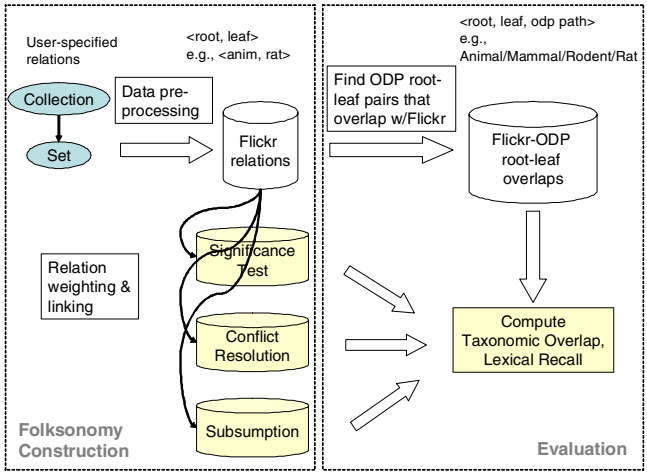


Figure 3: Overall architecture of the folksonomy learning and evaluation system

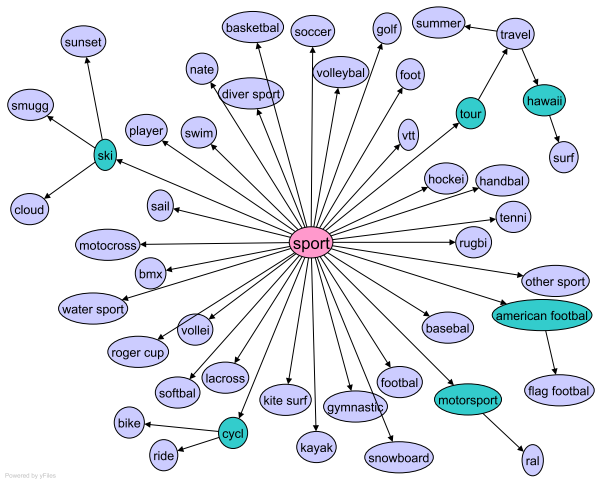


Figure 4: Folksonomy associated with concept `sport`.

the term co-occurrence between a and b is computed from the number of all bags-of-terms in which these two terms appear together. These two numbers are used to compute conditional probabilities $p(a|b)$ and $p(b|a)$. Then, a subsumes b if and only if $p(a|b) \geq t$ and $p(b|a) < t$, where t is an adjustable threshold, which can be determined empirically.[5] After all subsumption relations are found, we link them together and use the longest path as the path selection criterion, as described in Section 3.2.2.

4. EMPIRICAL RESULTS

For our study, we gathered data about collection/set relations created by a subset of Flickr users. To gather list of users, we used the Flickr API to retrieve the names of members of seventeen public groups devoted to wildlife and naturale photography. We then used a Web page scraping tool to retrieve collection and set hierarchies created by these users. Of the 39,922 users in our set, 21,792 created at least one collection, and about 600 users created multi-level, or collections of collections, hierarchies. The subjects covered by users' photographs were broad ranging, but a few common themes emerged. In addition to wildlife and nature photography, other common subjects were travel and sports photography, arts and crafts, and people and portraiture. We then used the methods described in this paper to aggregate many independently created shallow hierarchies into a common deeper folksonomy.

The architecture of our folksonomy learning system is shown in Figure 3. After preprocessing data, we obtained 215,537 relations, with 102,259 unique concept names. Subsequently, these relations are fed to the different relation weighting and linking schemes. After filtered relations are linked into deeper folksonomies, we first qualitatively investigate them using yEd graph editor[6], and then compare the learned folksonomies to reference taxonomies.

[5]We use a variant version of [16] proposed by Schmitz [17] although these two versions have negligible differences in our empirical studies.

[6]http://www.yworks.com/en/products_yed_about.html

4.1 Qualitative Evaluation

The resulting graph of interlinked concepts is quite complex. To simplify browsing, we extract subgraphs associated with a concept. Starting with a given root concept, we follow outgoing relations on the graph to get the children (narrower concepts) and their children, etc, four levels deep. We illustrate here the results with sample graphs, constructed using significance approach with $\alpha = 0.01$. The graph in Figure 5 shows the concept graph for the (stemmed) `country`. Its children include `france`, `china`, `india`, `uk`, etc. All of the children of `country` are proper countries. The child concepts of individual countries correspond to cities or landmarks within those countries. For example, `russia` has narrower concepts `moscow`, `st petersburg`, and `hermitage`, while `usa` has `new england` as one of its children, which itself has `massachusetts` and `connecticut` as children, with `massachusetts` also the parent of `cape cod`. In general, the automatically discovered concepts are quite useful, although not perfect. The algorithm does not distinguish between granularity levels of different concepts. For `usa`, for example, states, cities, and national parks are added at the same level. In addition, `united states` is a separate node, with a few of its own children, such as `texas`.

While geographical names provide a common vocabulary for labeling and organizing travel photographs, there is sufficient vocabulary commonality to induce folksonomies in other domains. We present three more folksonomies to illustrate our method's ability to discover many relevant sub-concepts. Figure 6 shows the graphs associated with (a) `invertebrate` and (b) `vertebrate`. The `vertebrate` folksonomy includes `bird` and many specific types of birds, reflecting the fact that bird watching is a passion of many avid naturalists armed with cameras. Our method discovered many useful sub-concepts of `invertebrate`, but put `moth` as narrower concept of `spider`, which is not correct. The `sport` folksonomy in Figure 4 shows many specific types of sports. However, our algorithm incorrectly associated `ski` with `cloud` and `sunset`, because `skiing` and `sky` both stem to `ski`.

Compared to folksonomies learned by *Significance Test* approach, those learned by the *Conflict Resolution* method

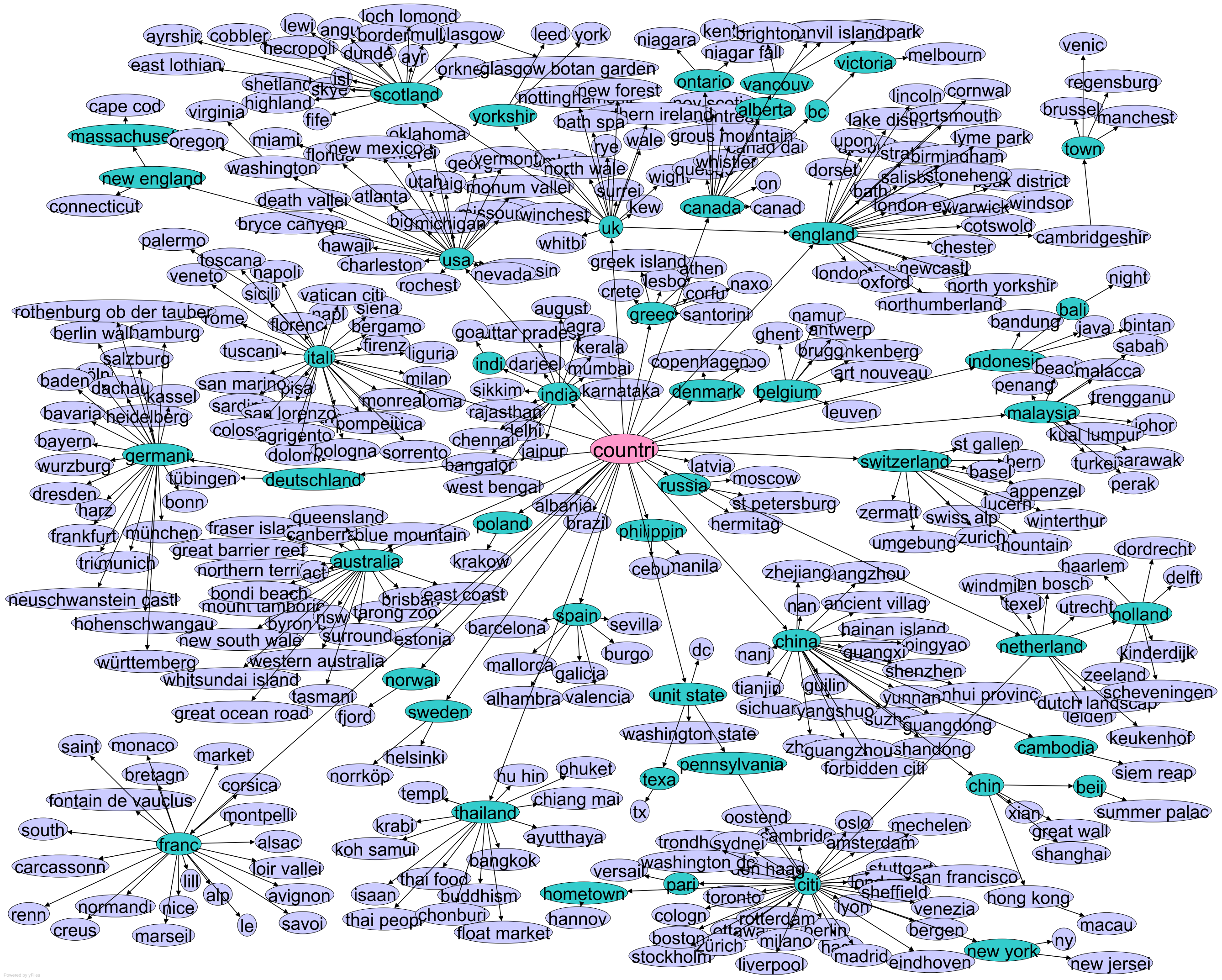


Figure 5: Folksonomy associated with the concept `country`, with root concept in pink.

are more densely linked, sometimes to irrelevant concepts, while *Term Subsumption* induces much shallower folksonomies, where many informative concepts are ignored. We provide quantitative comparison among the three approaches in the next section.

4.2 Quantitative Evaluation

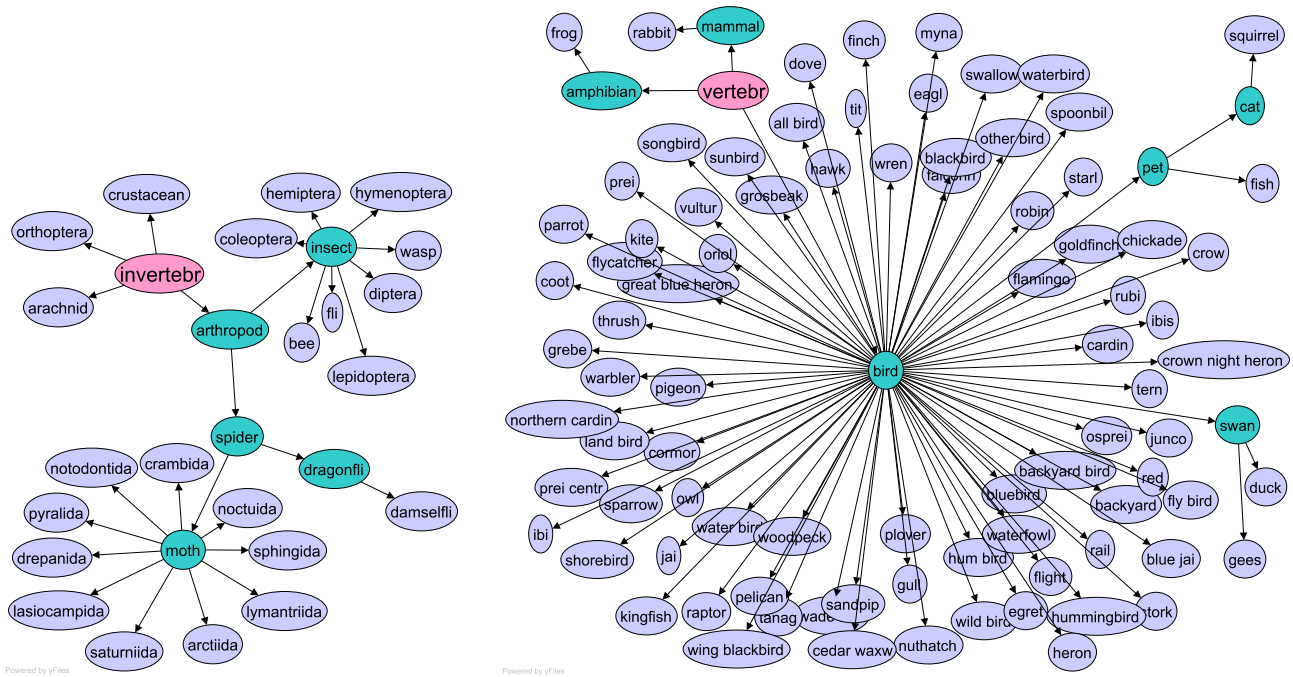
In this section, we describe methodology to quantitatively evaluate the quality of the learned folksonomies. Instead of asking human subjects to assess folksonomies' quality, we automatically evaluate them by comparing them to existing hand-built taxonomies. We first describe the overall process of the evaluation, shown in Figure 3, and then the metrics we use.

4.2.1 Approach

Human judgement was used in many previous works on automatic ontology construction, e.g. [16, 19, 17], to measure quality of induced ontologies. Although such evaluation is very natural, performing unbiased assessment on a huge collection of taxonomies is, however, an extremely expensive and time-consuming task. As hand-crafted taxonomies such as WordNet and Open Directory Project become freely available, one possible alternative is to compare how "similar" the induced folksonomies are to hand-crafted taxonomies.

One has to take into account at least two issues when comparing two taxonomies: (1) how many concepts are shared between the two taxonomies (scope), which will make the comparison meaningful, and (2) what similarity measure to use (metric). In response to the first issue, we propose using taxonomies from Open Directory Project (ODP).[7] The main reason we selected ODP is that, in contrast to WordNet, ODP is generated, reviewed and revised by many registered users. These users seem to use more colloquial terms than those used in WordNet. In addition, like Flickr users, they specify less formal relations, mainly broader/narrower relations. WordNet, on the other hand, specifies a number of formal relations among concepts, including hypernymy and meronymy. Note that ODP provides an alternative paradigm for community knowledge creation — rather than

[7]http://rdf.dmoz.org/



(a) (b)

Figure 6: Folksonomies associated with the concepts (a) `invertebrate` and (b) `vertebrate`.

synthesize knowledge from pieces of information created independently by many users, ODP (like Wikipedia) allows a large number of users work on a single document. Although any user can register to become an editor, she has to learn the structure and vocabulary and abide by ODP rules.

After we tokenized and stemmed ODP terms following the steps outlined in Section 3.1, we found 166,153 unique terms (*cf* 110,543 unique Flickr terms) with 15,495 terms in common. This proportion demonstrates that Flickr concepts somewhat overlap ODP in scope. Comparing the entire ODP data to our learned folkonomies is impractical, since there is a very large number of possible subtrees that can be compared. We simplify this task by selecting a concept that exists in both ODP and Flickr (picked manually or randomly), and then treat it as a root of the tree for each data set. We span the tree from the root concept. The depth of the tree is not imposed directly. Instead, we use the following methodology to pick "leaf" concepts of the tree in the Flickr data. In the Flickr relations set, we start at a specific "root" concept. We span the tree following relations for a given number of hops. We use only two spanning hops because Flickr concepts are densely linked. All concepts at which the spanning hops terminate — either because they have no children, or the number of hops has reached maximum – are then chosen as leaf candidates.

Once we have a specific root concept and a set of leaves, we select a tree from the learned Flickr folksonomy that covers these concepts, and also a tree from the ODP that covers these concepts. The two trees are then compared using the metrics described in below. Note that some leaves may not appear in the selected folksonomies since they are filtered out by the relation weighting schemes. Meanwhile, they may not appear in the selected ODP taxonomies due to a difference in scope between ODP and Flickr.

4.2.2 Metrics

Maedche and Staab [10] proposed a method to measure similarity between two taxonomies. In this paper, we applied two of their measures: *Lexical Recall* and *Taxonomic Overlap* to measure if the learned folksonomies are similar to taxonomies in the ODP.

According to [10], *Lexical Recall* measures how well a taxonomy induction process can discover concepts that exist in the actual taxonomy, regardless of the correctness of the structure of the learned taxonomy. For simplicity, we also ignore polysemy issue, i.e., we assume that concepts with the same name are the same. Let C_1 be a set of all concepts in the learned taxonomy T_1, and let C_2 be the set of concepts in the reference taxonomy T_2. Lexical Recall is defined as $LR(T_1, T_2) = \frac{|C_1 \cap C_2|}{|C_2|}$.

Taxonomic Overlap is a similarity measure that takes into consideration taxonomy structure. In particular, each concept in a learned taxonomy and a corresponding concept in a reference taxonomy are compared on how much their ancestors and descendants overlap. A set of super-concepts (ancestors) and sub-concepts (descendants) of a given concept c in a taxonomy T is referred to as Semantic Cotopy (SC), which is defined according to [10] as:

$$SC(c, T) := \{c_j \in T | c <_T c_j \cup c >_T c_j\}. \quad (2)$$

Note for (2) that $c <_T c_j$ returns all descendants of c in taxonomy T, and $c >_T c_j$ returns ancestors of c. Unlike in the original formulation of SC, we do not include the node c to avoid overly optimistic evaluation.

Taxonomic Overlap ($\overline{TO}$) between two taxonomies can be determined from the average of degree of overlap between SCs of concepts in two taxonomies. According to [10], the $\overline{TO}$ of taxonomy T_1 and T_2 is:

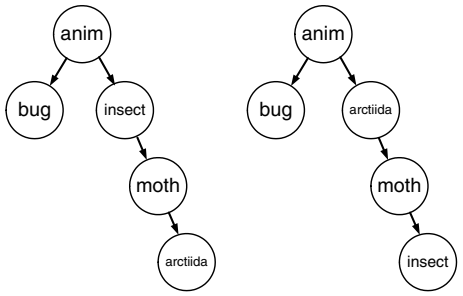


Figure 7: Illustrations of (a) a correct tree about "moth", and (b) an incorrect version of (a) where "insect" and "arctiida" (arctiidae) are misplaced. Original $\overline{TO}$ will judge the trees identical.

$$\overline{TO}(T_1, T_2) = \frac{1}{|C_1|} \sum_{c \in C_1} TO(c, T_1, T_2) \quad (3)$$

where

$$TO(c, T_1, T_2) := \begin{cases} TO^{'}(c, T_1, T_2) \text{ if } c \in C_2 \\ TO^{''}(c, T_1, T_2) \text{ if } c \notin C_2 \end{cases}, \quad (4)$$

and where $TO^{'}$ and $TO^{''}$ are defined as:

$$TO^{'}(c, T_1, T_2) := \frac{|SC(c, T_1) \cap SC(c, T_2)|}{|SC(c, T_1) \cup SC(c, T_2)|} \quad (5)$$

$$TO^{''}(c, T_1, T_2) := \max_{c' \in C_2} \frac{|SC(c, T_1) \cap SC(c', T_2)|}{|SC(c, T_1) \cup SC(c', T_2)|} \quad (6)$$

Note that (6) makes an optimistic assessment when a concept name c in T_1 does not exist in T_2 by picking c' in T_2 that yields the best SC match to c in T_1. In other words, the method assumes that c' refers to the same concept as c, although their names are different.

We discovered that the original version of $\overline{TO}$ (3) does not penalize for incorrect concept ordering. Consider two trees in Figure 7. Since SC of "insect", "moth" and "arctiida" are the same for both trees, $\overline{TO}$ in (3) will judge trees (a) and (b) to be identical ($\overline{TO} = 1.0$). This is because SC in (2) considers all ancestors and descendants, regardless of their ordering. One possible solution is to consider concept's ancestors and descendants *separately*. We modify (3) as follows:

$$\overline{TO}(T_1, T_2)^* = \frac{1}{2} \cdot \Big(\frac{1}{|C_1^{-root}|} \sum_{c \in C_1^{-root}} \hat{TO}(c, T_1, T_2) + \frac{1}{|C_1^{-leaves}|} \sum_{c \in C_1^{-leaves}} \check{TO}(c, T_1, T_2) \Big) \quad (7)$$

where C_1^{-root} is a set of all concepts in T_1 *except* its root concept. We exclude the root concept because it has no ancestors. Similarly, $C_1^{-leaves}$ is a set of all concepts in T_1 *except* its leaf concepts. $\hat{TO}$ ($\check{TO}$) is computed as in (5), but uses $\hat{SC}$ ($\check{SC}$) instead of SC. We define $\hat{SC}$ as *ancestor* Semantic Cotopy, which only considers ancestors of a certain concept, and $\check{SC}$ as *descendant* Semantic Cotopy, which only considers descendants of the concept:

$$\hat{SC}(c, T) := \{c_j \in T | c >_T c_j\}, \quad (8)$$

$$\check{SC}(c, T) := \{c_j \in T | c <_T c_j\}. \quad (9)$$

Returning to the case in Figure 7, the modified TO metric can detect that trees (a) and (b) have different concept ordering: $\overline{TO}(T_b, T_a)^* = 0.417$. Since $\overline{TO}$ is not symmetric as pointed out in [4], one can compute a harmonic mean between $\overline{TO}(T_1, T2)$ and $\overline{TO}(T_2, T1)$ to get a symmetric score.

Another measure we also use is "average path depth". Basically, this measure gauges an average depth of paths from root to all leaf nodes in a given taxonomy. A depth of a certain path is a number of hops (or relations) in the path. The average path depth of Figure 7 (b) is then $\frac{(1+3)}{2} = 2$.

4.2.3 Quantitative Comparison

Table 1 presents performance of proposed approaches: *Conflict Resolution* and *Significance Test*, and a baseline approach *Term Subsumption* on 3 different metrics, comprising of *modified Taxonomic Overlap*, *Lexical Recall* and average depth of paths from root to leaves. We manually selected 32 different root concepts and use the methodology previously described in Section 4.2.1. These root concepts are about living things, objects and locations, which are mostly used by users in Flickr to describe their photos. Note that in the experiment, we used $t = 0.6$ for *Term Subsumption* approach.[8] Since we directly select the root concepts, to avoid biased comparison, we also modify (8) to exclude the root node of T.

As revealed by *modified Taxonomic Overlap*, folksonomies induced by *Conflict Resolution* and *Significance Test* are more consistent with corresponding ODP taxonomies than those induced by *Term Subsumption* approach. In most cases, *Significance Test* is somewhat superior to *Conflict Resolution*. Nevertheless, there is one case, "south africa", that *Term Subsumption* slightly performs better than the other twos. In such case, although all approaches induced about the same small number of concepts, *Conflict Resolution* and *Significance Test* induced one more concept, "kruger nate park", that does not exist in ODP.

In most cases, since *Term Subsumption* discards a greater number of informative concepts, comparing to *Conflict Resolution* and *Significance Test*, *Lexical Recall* of the former is much smaller. Furthermore, *Term Subsumption* induce much shallower folksonomies than those induced by the other twos. One reason why *Term Subsumption* discards many informative concepts and their relations in this context is that, a certain concept usually relates to many other concepts. Thus, it is very likely that a number of cooccurrences of a given concept pair is very low, compared to that of individual one. Consequently, a chance that one concept "subsumes" another one is very low. In our approaches, we instead consider explicit relations of concepts, which will not suffer from this issue.

5. RELATED WORK

Many researchers have studied the problem of constructing ontological relations from text, *e.g.*, [6, 14, 19]. These

[8]We tried different values for t between 0.8 to 0.55 and found that, at $t = 0.6$, the algorithm can induce folksonomies reasonably good, while not discarding too many concepts

	$\overline{fTO}^*$			Lexical Recal (LR)			Avg Path Depth			
Root Node	subs	conres	sig001	subs	conres	sig001	subs	conres	sig001	ODP
anim	0.0006	**0.0628**	0.0421	0.0128	**0.1848**	0.0821	1.01	1.99	2.19	3.29
bird	0.0087	**0.0302**	0.0032	0.0359	**0.1231**	0.0872	1.15	2.58	1.37	2.42
invertebr	-	0.1041	**0.1658**	0.0769	**0.3846**	**0.3846**	1.00	3.18	3.13	1.90
vertebr	-	0.0019	**0.0164**	-	0.2000	**0.3000**	-	2.39	2.06	1.83
insect	-	0.0022	**0.0033**	0.1429	**0.2857**	**0.2857**	1.06	2.17	1.28	1.40
fish	-	-	-	**0.0096**	-	-	1.00	-	-	3.10
plant	0.0010	0.0006	**0.0097**	0.0154	0.0308	**0.1154**	1.04	2.07	3.11	2.48
flora	-	0.0065	**0.0160**	0.0028	0.0850	**0.1530**	1.07	3.21	3.51	4.86
shrub	-	-	-	**0.0625**	-	-	1.00	-	-	2.67
fauna	-	0.0004	**0.0099**	0.0030	0.0151	**0.1118**	1.13	2.27	3.06	4.93
floral	-	0.0010	**0.0033**	-	**0.5000**	**0.5000**	-	2.24	3.30	1.00
flower	0.0000	0.0011	**0.0132**	0.0488	**0.0741**	0.0617	1.02	2.19	1.91	2.86
reptil	0.0095	**0.0740**	0.0619	0.1333	0.2000	**0.2667**	1.00	3.00	3.00	2.11
amphibian	-	**0.1687**	0.0062	-	**0.2083**	0.0833	-	2.75	1.00	1.95
build	-	-	-	**0.5000**	**0.5000**	**0.5000**	1.22	1.81	2.69	0.33
urban	-	**0.0015**	-	0.0323	**0.0645**	0.0323	1.00	2.74	2.28	2.38
countri	-	0.0146	**0.0188**	0.0101	**0.0808**	0.0505	1.00	2.29	2.47	2.07
africa	-	**0.1346**	0.1189	0.0062	**0.2099**	0.1173	1.00	2.12	1.37	3.01
asia	-	0.2260	**0.2406**	0.0018	**0.1871**	0.1646	1.00	2.69	2.32	3.30
europ	0.0002	0.1526	**0.1970**	0.0021	**0.1184**	0.1102	1.12	2.56	2.72	4.10
south africa	**0.0116**	0.0050	0.0050	**0.0385**	**0.0385**	**0.0385**	1.00	1.00	1.00	2.41
north america	-	**0.1030**	0.0880	-	**0.1013**	0.0953	-	2.98	3.18	5.02
south america	-	**0.2293**	**0.2293**	-	**0.1571**	**0.1571**	-	1.89	1.89	3.40
central america	-	**0.0927**	**0.0927**	-	**0.0667**	**0.0667**	-	2.00	2.00	3.44
unit kingdom	-	0.1343	**0.1389**	0.0012	**0.0753**	0.0718	1.00	3.22	3.01	3.46
unit state	-	**0.1023**	0.0866	0.0009	**0.0810**	0.0749	1.06	2.81	2.78	4.22
world	0.0001	0.0296	**0.0387**	0.0005	0.0432	**0.0439**	1.00	2.47	2.81	6.26
citi	-	0.0033	**0.0077**	0.0286	**0.1429**	0.0857	1.00	2.56	1.84	1.07
craft	-	**0.0157**	0.0071	0.0061	**0.0848**	0.0364	1.17	2.67	1.97	2.66
dog	-	**0.0002**	-	0.0060	**0.0119**	0.0060	1.00	2.14	1.00	4.10
cat	-	**0.0036**	-	0.0097	**0.0291**	0.0097	1.00	2.06	1.00	3.95
sport	0.0008	0.0290	**0.0322**	0.0073	**0.0377**	0.0261	1.00	1.76	1.33	3.74

Table 1: This table presents empirical validation using 3 different metrics: *modifited Taxonomic Overlap* (averaged using Harmonic Mean), *Lexical Recall* and the average depth of paths from root to all leaves. The scale for *modifited Taxonomic Overlap* and *Lexical Recall* is from 0.0 to 1.0 (the higher the better). Each folksonomy tree is represented by its root name as in the first column in each row. The column, named "subs", presents the performance of *Subsumption* for each folsonomy tree, as "conres" and "sig001" presents that of *Conflict Resolution*, *Significance Tests*with confidence level 0.01 respectively. As the last column, ODP, shows average depth of paths from root to leaves in Open Directory Project. In some cases, "-" exists because a corresponding approach does not induce any concept.

works exploit linguistic patterns to infer if two keywords are related under a certain relationship. For instance, they use "such as" to learn hyponym relations. Cimiano *et al.* [4] also applies linguistic patterns to extract object properties and then uses Formal Concept Analysis (FCA) to infer conceptual hierarchies. In FCA, a given object consists of a set of attributes and some attributes are common to a subset of objects. A concept 'A' subsumes concept 'B' if all objects in 'B' (with some common attributes) are also in 'A'. However, these approaches are not applicable to the metadata on social Web sites such as tags, bundles and photo sets, which are ungrammatical and unstructured.

Recently, several works proposed different approaches to construct conceptual hierarchies from the metadata collated from social Web sites. Mika [13] uses a graph based approach to construct a network of related tags, projected from either a user-tag or object-tag association graphs. Although there is no evaluation on inducing broader/narrower relations, the work suggests inferring them by using betweenness centrality and set theory. Other works apply clustering techniques to keywords expressed in tags, and use their co-occurrence statistics to produce conceptual hierarchies [3]. In a variation of the clustering approach, Heymann and Garcia-Molina [7] uses graph centrality in similarity graph of tags. In particular, the tag with the highest centrality would be more abstract than that with a lower centrality; thus it should be merged to the hierarchy before the latter, to guarantee that more abstract node gets closer to the root node. Schmitz [17] has applied a statistical subsumption model [16] to induce hierarchical relations of tags.

We believe that the previously mentioned works suffer from the "popularity vs generality" problem that arises when using tags to induce a hierarchy. Specifically, a certain tag may be used more frequently not only because it is more general, but because it is more popular among users. On Flickr, we found that there are ten times as many photos tagged with "car" than with "automobile." If we apply clustering approaches, "car" may be found to be more abstract than "automobile" since, the former is likely to have higher centrality than the latter. And if we apply statistical subsumption model, the former would be likely to subsume the latter since there is a higher chance that photos tagged with

"car" are also tagged with "automobile". Of course, we believe that tag statistics are a good source of evidence for inducing hierarchies; however, tag statistics alone may not be enough to discover conceptual hierarchies.

There is another line of research that focuses on exploiting partial hierarchies contributed by users. *GiveALink* project collects bookmarks donated by users [11]. Each bookmark is organized in a tree structure as folder and sub folders by an individual user. Based on tree structures, similarities between URLs are computed and used for URL recommendation and ranking. Although this project does not concentrate on conceptual hierarchy construction, it provides a good motivation to exploit explicit partial structures like folder and subfolder relations. Our approach is in the same spirit as *GiveALink* — we exploit collection and set relations contributed by users on a social Web site to construct conceptual hierarchies. We hypothesize that generality-popularity problem of keywords in collection-set relation space is less than that in tag space. Although people may use a keyword "car" far more than "automobile" to name their collections and sets, not so many people would put their "automobile" album into "car" super album.

Our approach is also similar in spirit to several works on ontology alignment (e.g. [21]). However, unlike those works, which merge a small number of deep and detailed concepts, we merge large number of noisy and shallow concepts, which are specified by different users.

6. CONCLUSION

The social Web sites allow users to contribute content and also provide tools to help them manage content by annotating it with descriptive tags, and more recently, with semantic relations. By making large amount of such metadata available, social Web sites enable researchers to empirically study how humans organize knowledge, and also to learn a common classification system, a folksonomy, from the data. This paper describes statistical approaches to aggregating large number of simple broader/narrower relations specified by different users into a common, deeper folksonomy. Empirical results desmonstrate that our approaches can induce quite detailed folksonomies, which are also more consistent with taxonomies in Open Directory Project than those produced by the previous approach. Our approach is general, and can be applied to other systems that allow users to specify relations: e.g., the social bookmarking site Del.icio.us allows users to group related tags into tag bundles.

Our long-term goal is to learn the structure of collective knowledge from the evidence provided by many users [8]. We believe that the simple relations described above are more informative than tags alone for learning how people classify things. In the future, we plan to separate "broader/narrower" from "related-to" relations. We also need to more systematically handle the challenges of different users using a different classification order and different level of specificity in the relations they specify. We would also like to combine relations with tag statistics to disambiguate concepts.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. CMMI-0753124 and IIS-0812677.

7. REFERENCES

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

[2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[3] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proc. of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM.

[4] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res. (JAIR)*, 24:305–339, 2005.

[5] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, April 2006.

[6] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. of ACL-92*, pages 539–545, Morristown, NJ, USA, 1992.

[7] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University, Stanford, CA, USA, April 2006.

[8] C. Kemp, A. Perfors, and J. B. Tenenbaum. Learning domain structures. In *Proc. of the 26th Annual Conference of the Cognitive Science Society*, 2005.

[9] K. Lerman. Social information processing in news aggregation. *IEEE Internet Computing: special issue on Social Search*, 11(6):16–28, November 2007.

[10] A. Maedche and S. Staab. Measuring similarity between ontologies. In *EKAW*, pages 251–263, 2002.

[11] B. Markines, L. Stoilova, and F. Menczer. Bookmark hierarchies and collaborative recommendation. In *Proc. of AAAI*, 2006.

[12] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[13] P. Mika. Ontologies are us: A unified model of social networks and semantics. *J. Web Sem.*, 5(1):5–15, 2007.

[14] M. Pasca. Acquisition of categorized named entities for web search. In *Proc. of the 13rd ACM international conference on Information and knowledge management*, pages 137–145, New York, NY, USA, 2004.

[15] A. Plangprasopchok and K. Lerman. Exploiting social annotation for automatic resource discovery. In *Proc. of AAAI workshop on Information Integration*, 2007.

[16] M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In *SIGIR*, pages 206–213, 1999.

[17] P. Schmitz. Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW Š06)*, May 2006.

[18] B. Shneiderman. Computer science: Science 2.0. *Science*, 319(5868):1349–1350, March 2008.

[19] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proc. of ACL-06*, pages 801–808, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[20] L. Steels and E. Tisselli. Social tagging in community memories. In *Proc. of AAAI symposium on Social Information Processing*. AAAI, 2008.

[21] O. Udrea, L. Getoor, and R. J. Miller. Leveraging data and structure in ontology integration. In *SIGMOD Conference*, pages 449–460, 2007.

[22] M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *ISWC/ASWC*, pages 680–693, 2007.