Constructing Folksonomies by Integrating Structured Metadata

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

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

Lise Getoor
Department of Computer Science
University of Maryland, College Park
getoor@cs.umd.edu

ABSTRACT

Aggregating many personal hierarchies into a common taxonomy, also known as a folksonomy, presents several challenges due to its sparseness, ambiguity, noise, and inconsistency. We describe an approach to folksonomy learning based on relational clustering that addresses these challenges by exploiting structured metadata contained in personal hierarchies. Our approach clusters similar hierarchies using their structure and tag statistics, then incrementally weaves them into a deeper, bushier tree. We study folksonomy learning using social metadata extracted from the photo-sharing site Flickr. We evaluate the learned folksonomy quantitatively by automatically comparing it to a reference taxonomy created by the Open Directory Project. Our empirical results suggest that the proposed approach improves upon the state-of-the-art folksonomy learning method.

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

Keywords

Folksonomies, Collective Knowledge, Data Mining

1. INTRODUCTION

Many social Web sites allow users to annotate the content with descriptive metadata, and to organize content hierarchically. These types of structured metadata provide valuable evidence for learning how a community organizes knowledge. Although these types of social metadata lack formal structure, they capture the collective knowledge of Social Web users. Once extracted from the traces left by many users, such collective knowledge will add a rich semantic layer to the content of the Social Web that will potentially support many tasks in information discovery, personalization, and information management.

Learning a global folksonomy comes with a number of challenges which arise when integrating structured metadata created by diverse users, with each user freely annotating data according to her own preferences. Consequently, social metadata is *noisy*, *shallow*, *sparse*, *ambiguous*, *conflicting*, *multi-faceted*, and expressed at *inconsistent granularity levels* across many users.

Previous works, e.g., [2, 4], addressed some of the above challenges. Basically, they induce folksonomies from tags by utilizing tag statistics. They assume that more frequent tags describe more general concepts. However, as pointed out in the late work [3], frequency-based methods cannot distinguish between more general and more popular concepts. [3], on the other hand, overcame the "popularity vs generality" problem by using user-specified relations extracted from personal hierarchies. This method addressed the challenge of conflicting metadata by keeping relations that many users agreed on, but it did not exploit tag statistics and structure information, which can potentially resolve the ambiguity challenge.

2. STRUCTURED SOCIAL METADATA

In addition to keywords or tags to describe content, some social Web sites also allow users to organize content hierarchically. In *Flickr*, for instance, users can *arbitrarily* group related photos into *sets* and then group related sets in *collections*. Some users create multi-level hierarchies containing collections of collections, etc., but the vast majority of users who use collections create shallow hierarchies, consisting of collections and their constituent sets. These personal hierarchies generally represents subclass and part-of relationships. Even without strict semantics being attached to these relations, we believe that these personal hierarchies represent a novel, rich source of evidence for learning folksonomies, which express how a certain community organizes contents.

3. LEARNING FOLKSONOMIES FROM STRUCTURED METADATA

We propose a simple, yet effective approach to combine many personal hierarchies into a global folksonomy. We first define a personal hierarchy as a shallow tree, a *sapling*, composed of a root node r^i and its child, or leaf, nodes $\langle l_1^i, ..l_j^i \rangle$. The root node corresponds to a user's collection, and inherits its name, while the leaf nodes correspond to the collection's constituent sets and inherit their names. We assume that hierarchical relations between a root and its children, $r^i \rightarrow l_j^i$, specify broader-narrower relations.

In addition to hierarchical structure, each sapling carries information derived from tags. On Flickr, users can attach tags only to photos. A sapling's leaf node correspond to a set of photos, and the tag statistics of the leaf are aggregated from that set's constituent photos. Tag statistics are then

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

propagated from leaf nodes to the parent node. We define a tag statistic of node x as $\tau_x := \{(t_1, f_{t_1}), (t_2, f_{t_2}), \cdots (t_k, f_{t_k})\}$, where t_k and f_{t_k} are *tag* and its frequency respectively. Hence, τ_{r^i} is aggregated from all $\tau_{l^i_j}$s.

3.1 Relational Clustering of Structured Metadata

In order to learn a common folksonomy, we need to aggregate saplings *both* horizontally and vertically. By horizontal aggregation, we mean merging saplings with similar roots, which expands the breadth of the learned tree by adding leaves to the root. By vertical aggregation, we mean merging one sapling's leaf to the root of another, extending the depth of the learned tree. The approach we use exploits contextual information from neighbors in addition to local features to determine which saplings to merge. The approach is similar to relational clustering[1] and its basic element is the similarity measure between a pair of nodes.

We define a similarity measure between nodes in different saplings, which combines heterogeneous evidence available in the structured social metadata, and is a combination of *local similarity* and *structural similarity.* The local similarity between two nodes a and b, $localSim(a, b)$, is based on the intrinsic features of a and b, such as their names and tag distributions. The structural similarity, $structSim(a, b)$ is based on features of neighboring nodes. If a is a root of a certain sapling, its neighboring nodes are all of its children. If a is a leaf node, the neighboring nodes are its parent and siblings. The similarity between nodes a and b is:

$$\begin{aligned} nodesim(a, b) &= \alpha \times localSim(a, b) \qquad (1) \\ &+ (1 - \alpha) \times structSim(a, b), \end{aligned}$$

where $0 \leq \alpha \leq 1$ is a weight for adjusting contributions from $localSim(,)$ and $structSim(,)$.

SAP: Growing a Tree by Merging Saplings

Our algorithm which uses the similarity function defined above to incrementally grow a deeper, bushier tree by merging saplings created by different users. In order to learn a folksonomy corresponding to some concept, we start by providing a seed term, the name of that concept. The seed term will be the root of the learned tree. We cluster individual saplings whose roots have the same name as the seed by using the similarity measure to identify similar saplings. Saplings within the same cluster are merged into a bigger sapling. Each merged sapling corresponds to a different sense of the seed term.

Next, we select one of the merged saplings as the starting point for growing the folksonomy for that concept. For each leaf of the initial sapling, we use the leaf name to retrieve all other saplings whose roots are similar to the name. We then merge saplings corresponding to different senses of this term as described above. The merged sapling whose root is most similar to the leaf is then linked to the leaf. In the case that several saplings match the leaf, we merge all of them together before linking. Clustering saplings into different senses, and then merging relevant saplings to the leaves of the tree proceeds incrementally until some threshold is reached.

4. EMPIRICAL VALIDATION

We used the data set containing collections and their constituent sets (or collections) created by a subset of Flickr users [3]. We compare SAP against the folksonomy learning method, SIG, described in [3]. Briefly, SIG first breaks a given sapling into (collection-set) relations. The approach then employs hypothesis testing to identify the most informative relations. Informative relations are then linked into a deeper folksonomy. We used a significance test threshold of 0.01.

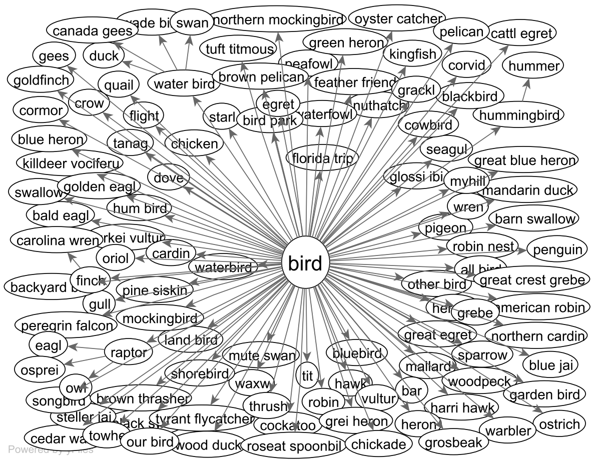

Figure 1: Folksonomies learned for bird

We quantitatively evaluate the induced folksonomies by automatically comparing them to the reference hierarchies extracted from the Open Directory Project(ODP). We first manually select 34 seed terms to induce folksonomies by SAP and SIG We then use methodology described in [3] to automatically evaluate on how consistent they are, with respect to the reference hierarchies. Generally, in 75% of the cases, SAP produced bushier trees; and recovers a larger number of concepts, relative to ODP, as indicated by the numbers of overlapping leaves (in 90% of the cases) and better Lexical Recall scores (in 62.5% of the cases). In addition, SAP can produce trees, which are structurally consistent to the ODP, as indicated by Taxonomic Overlap score (in 77% of the cases).

After closely inspecting the learned trees, we found that SAP demonstrates its advantage over the baseline in disambiguating and correctly attaching relevant saplings to appropriate induced trees. For instance, `bird` tree produced by SAP does not includes `Istanbul` or other Turkey locations, as shown in Figure 1.

In all, the proposed approach has several advantages over baseline. First, it cautiously combines relevant saplings, based on contextual evidence, which can resolve ambiguity of the concept names. Second, only a seed is required to incrementally build a tree, while both seed and leaf nodes are required by the SIG method. Third, it allows similar concepts to appear multiple times within the same hierarchy.

5. REFERENCES

[1] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 1(1):5, 2007.

[2] 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.

[3] A. Plangprasopchok and K. Lerman. Constructing folksonomies from user-specified relations on flickr. In *WWW*, 2009.

[4] P. Schmitz. Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW '06)*, May 2006.