On constructing shallow taxonomies from social annotations

Anon Plangprasopchok and Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
{plangpra,lerman}@isi.edu

Abstract

Tagging in social media system has demonstrated to be a convenient way for users to annotate objects of interest.One reason behind its success obviously because tags can be chosen by users arbitrarily without any topic and specificity constraints. Although tags are free-from keywords, there are some evidences [1] suggesting that, for a particular object type, users tend to use "similar" tag sets. In addition, such tags are in different levels of specificity. This might suggest that there are some hierarchical concepts behind users' tagging processes. In this paper, we outline a problem in extracting hierarchical concepts from social annotation data and propose a possible solution – a probabilistic generative model that describes tagging processes.

Introduction

As users continue to flock to social media sites, e.g., the social photo-sharing site *Flickr* and social bookmark-sharing site *Del.icio.us*, they generate an ever increasing quantity of content and metadata. The metadata comes in different formats: users discuss content and label it with freely chosen keywords, known as *tags*. Tags were originally meant to make it easier for users to manage their won content, e.g., to easily find and retrieve relevant documents. There are few, if any, constraints placed on tags. A user is free to choose the scope, specificity, and even semantics of the tag are idiosyncratic to the user.

Although tags are generated arbitrarily, there is evidences suggesting that users tends to use a similar tags to label similar types of objects. Golder and Huberman (Golder & Huberman 2006) found that after about 100 users tagged a specific site, the distribution of tags for that site remains constant, meaning that no new tags are introduced by subsequent users. A geocoding Web site that provides coordinates of a given address is tagged on del.icio.us with keywords "geocoding", "gis","latitude" and "maps", which are closely related in some way. Similarly, Flickr photos submitted to the same group are generally tagged with the same set of keywords. Most Flickr users in "insect macro group", for instance, use related keywords like "insect", "flower", "bee" and " butterfly" to tag photos in this group. In addition, without any modification on users' tags, our previous works (Lerman, Plangprasopchok, & Wong 2007; Plangprasopchok & Lerman 2007) and Flickr team [2] demonstrates some emergences of semantic clusters by analyzing tag occurrences, generated by multiple users. These evidences may suggest that people have similar concepts or some common agreement to categorize an object; and thus we can perceive similar objects tagged by similar tags across users.

Users tend to provide tags at multiple levels of specificity. Thus, for example, we find users who use more abstract tags like "geocoding" along with more specific tags like "latitude" to describe a geocoding Web site. In Flickr photos, tags "Bee" and "butterflies" may often come with tags "insect" or "bugs". We want to exploit this evidence to automatically discover taxonomy of concepts inherent in socially generated metadata. Such collectively created taxonomies, referred to as folksonomies, will facilitate information management by, e.g., making it easier to find content relevant to user's query or integrate information from disparate sources. Researchers have recently began to address the problem of discovering latent hierarchies present in social annotation (Mika 2007). By using a subsumption-based model, derived from co-occurrences of tags along with appropriate configurations, (Schmitz 2006) has showed that such model can induce taxonomies from flat tags. In an unpublished work, we used a probabilistic model to recover tags that are more general than the others, by tweaking the model to have one latent topic to be equally likely selected to tag a certain set of objects.

In this paper, we describe a probabilistic model that can infer hierarchies of concepts from social annotation. The model exploits the fact that different users tag objects at different levels of specificity (cite basic level theory). In addition, it exploits evidence from the explicit relations that several social Web sites allow users to create. Del.icio.us, for example, allows users to *bundle* related tags, while Flickr users can group related images in *sets* (or albums), and related sets in *collections*. Below we describe this evidence

[1] most common tags in del.icio.us pages demonstrates that relevant tags always chosen by users to annotate a certain URL

[2] one feature in Flickr "clusters" allow user to find related photos and tags given a certain tag. http://flickr.com/photos/tags/lion/clusters/ shows clusters of tags and photos related to "lion"

and our approach to inferring hierarchies of concepts from it.

Learning Taxonomies

In the past, researchers have attempted to discover taxonomies, or topic hierarchies, in a collection of documents using simply the distribution of words across documents (Mimno, Li, & McCallum 2007). We can apply a similar heuristic to discovery taxonomies through tagging (Schmitz 2006): more general tags tend to appear more frequently than more specific tags on the "same kinds" of objects. Thus, to generate a precise taxonomies using such heuristic, one must have a good object partition that contains objects of the "same kind" [3]. In many cases, we may not know such a partition a priori, and a criterion to decide if a partition is valid is difficult to obtain.

Another important issue that affects social annotation, and therefore the learning of taxonomies, comes from variations tag usage, that includes a variation in basic-level categories. A user may choose to highlight different facets of the object. For example, one Flickr user may always add a tag related to the place where she took a photo; while another may add a tag related to equipment she used for taking the photo. In addition, a given item can be described by terms along a spectrum of specificity, ranging from specific to general. A Siberian tiger can be described as a "tiger," but also as a "mammal" and "animal." The basic level is the category people choose for an object when communicating to others about it. Thus, for most people, the basic level for canines is "dog," not the more general "animal" or the more specific "beagle." However, what constitutes the basic level varies between individuals, and to a large extent depends on the degree of expertise. To a dog expert, the basic level may be the more specific "beagle" or "poodle," rather than "dog." The basic level problem arises when different users choose to describe the item at different levels of specificity. For example, a dog expert tags an image of a beagle as "beagle," whereas the average user may tag a similar image as "dog." We claim that a simple approach, which simply aggregates tag occurrences across all users and then determines a cutoff for tag clusters and specificities, may suffer from such variations.

In addition to statistical properties of tags, there is other important evidence useful for learning taxonomies. On Flickr for example, a user can group pictures of damselflies and dragonflies together in a set (or an album) called "Dragonflies", and group the sets "Dragonflies", "Beetles", "Spiders", etc., together in a collection called "Insects". The collections can be thought of as specifying a subsumption relation on sets, while sets provide a subsumption relation on tags of the included images. Similarly, tag bundles in del.icio.us group related tags together and provide a subsumption relation across them. Exploiting this evidence can potentially improve the quality of extracted taxonomies.

[3]This is quite obvious in Flickr: since usually only owners can tag their photos, one might have to cluster the "same" photos together before constructing taxonomies

We propose a probabilistic generative model that describes users' tagging process which addresses the issues mentioned so far. The assumptions we make for the model are as follows: (1) there are common taxonomies of concepts shared among users; (2) according to object categories, user interests and specificities, tags are generated from such trees; (3) it is possible to have some variations of specificities across users but one user must have the same specificity in tagging any object.

Problem Definition

Given a set of observations of annotated objects O, $\langle u, \{t..\}, \{s..\}\rangle_o$, each of which is composed of user u, tag t and object (or tag) set s, identify super and sub topics, z^s and z respectively, to which some tags associate, as well as compute dependencies among z^s and z. The topics z^s and z are defined as categorical concepts; and both of them are hidden variables, whose values have to be estimated in model's parameter estimation process.

Proposed Model

We propose to extend the hierarchical Pachinko Allocation Model (Mimno, Li, & McCallum 2007) to extract taxonomies in social annotation context. We postulate that an object is annotated with tags coming from a variety of latent topics at different levels of specificity. We also postulate that we can separate these topics into two sets, with one set of topics acting as a super-category (super-topic) of the other. When a user creates a tag to annotate an object, a path is chosen through a taxonomy, traversing super- and sub-topic categories. Based on user's preferred specificity level, one among super- or sub- topics in the chosen path is selected; then a tag associated with that topic is selected for the object.

We will specifically examine the case where some objects have categorical constraints imposed by users, such as sets and collections on Flickr, or tag bundles on Delicious. Such constraints can be incorporated in the model by using them as observations of super-topics. That is, after choosing an object's topic path, a user assigns the object to a set, corresponding to the super-topic of that path. We will study the model within the Flickr environment.

Formally, a process of annotating objects in a social annotation system can be described as follows.

- For each object-user annotation, $\langle u, \{t..\}, \{s..\}\rangle_o$, sample a distribution θ^s over super-topics. Subsequently, for each super-topic, sample a distribution θ over sub-topics.
- For each tag t in annotation o
  - sample super-topic z^s from θ^s
  - sample sub-topic z from θ_{z^s}
  - sample specificity l from λ_u
  - sample tag t from z^s or z according to l
- For each set s in annotation o
  - sample super topic z^s from θ^s
  - sample set s from σ_{z^s}

With appropriate prior distributions (hyperparameters) on θ^s, θ, ϕ, σ and λ, one can analytically integrate them out,

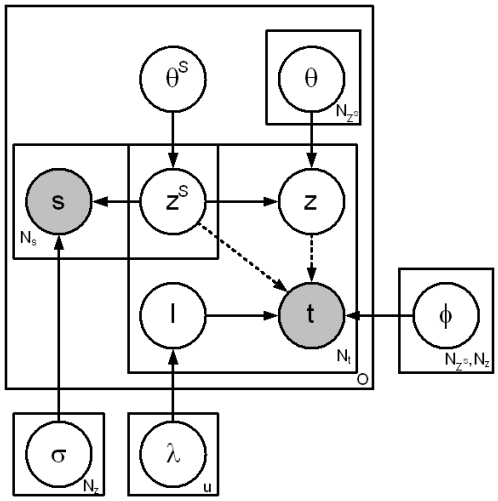

Figure 1: Graphical representation of the probabilistic generative model for extracting taxonomies from social annotation. t, u, s, l, z^s and z denote variables "tag", "user", "set", " level", "super-topic" and "sub-topic" respectively. N_t represents a number of tag occurrences for a object-user annotation; O represents a number of objects, N_z represents a number of sub topics, N_{z^s} represents a number of super-topics, and N_s represents a number of sets associated with the object. Dotted lines from z^s and z to t indicate that one of z^s and z will be chosen to generate t, according to specificity l. Note that filled circles represent observed variables. Hyperparameters are omitted for sake of simplicity.

and the remaining task is to estimate z^s, z and l for each object annotation. We plan to use Gibbs sampling for such estimation. Dependencies between super and sub topics can be obtained by computing $p(z|z^s)$ and those between topics and words by $p(w|z)$ (or ϕ)

Application to Flickr

We plan to apply the model above to learn the hidden topic hierarchy of topics (or concepts) from a collection of tagged images submitted to Flickr. We plan to aggregate images from a variety of groups of interest to amateur naturalists. These groups are places for avid naturalists to post photographs of insects, birds and flowers. We plan to compare the automatically discovered topic hierarchy with a formal taxonomy, such as the Linnaean classification system for these organisms.

Conclusion

This paper describes an approach to automatically discovering topic hierarchies (folksonomies) from socially annotated data on Web sites Flickr and Delicious. These sites are interesting because in addition to straight metadata such as tags, users are also allowed to specify constraints or relations between objects, for example, by grouping related objects together in a collection. We propose to extend a hierarchical probabilistic model to learn from evidence (tags and relations) in this domain. The model will take advantage of the natural variation in the levels of specificity which different users resort to while tagging content to derive the taxonomy of concepts. The automatically learned taxonomy is useful for a number of reasons. It can help organize the collectively contributed content, helping users browse through it to find relevant content. It can be used to integrate diverse content by enabling applications to automatically figure out the semantics of data (or at least whether within a taxonomy the data fits). It can also help shed light on the cognitive models shared by many users.

References

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

Lerman, K.; Plangprasopchok, A.; and Wong, C. 2007. Personalizing image search results on flickr. In *Proceedings of AAAI workshop on Intelligent Web Personalization*.

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

Mimno, D.; Li, W.; and McCallum, A. 2007. Mixtures of hierarchical topics with pachinko allocation. In *ICML*.

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

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