A Categorical Model for Discovering Latent Structure in Social Annotations

Said Kashoob and **James Caverlee**
Texas A&M University
Department of Computer Science

Ying Ding
Indiana University
School of Library and Information Science

Abstract

The advent of social tagging systems has enabled a new community-based view of the Web in which objects like images, videos, and Web pages are annotated by thousands of users. Understanding the emergent semantics inherent in the socially-generated collection of annotations has important research implications for information discovery and knowledge sharing. To this end, we propose a novel probabilistic generative model for discovering latent structure in large-scale social annotations. The generative model identifies latent community-based "categories" of interest that can be used to group semantically-related tags and to augment traditional content-based information search and discovery. We illustrate the proposed approach over large collections of Web objects annotated by the Flickr and Delicious communities. Additionally, we show how to integrate the annotation-based categorical model with traditional content-based approaches for the effective focused discovery and exploration of Web objects.

Introduction

The emerging Social Web is noted for wide-scale user participation in the generation, annotation, and sharing of information. In particular, the excitement surrounding social tagging systems – like CiteULike, Delicious, Flickr, and Newsvine, among many others – has been remarkable in the last few years, driving a growing interest in new avenues for information sharing and knowledge discovery

Social annotations (or tags) are typically simple keywords or phrases that can be attached to an object as informal user-specific metadata. For example, on the Delicious social tagging service, a user could tag the Web resource `www.espn.com` with tags like "sports", "my-favorites", and "scores". In isolation, a user's annotations can help organize a single user's bookmarks. But in the aggregate, the many tags applied by thousands of (largely) independent users can be used to uncover the collective intelligence (i.e., emergent semantics) for supporting smarter tag-based browsing (Bao et al. 2007), search (Li et al. 2007), and information access (e.g., through tag-based clustering (Brooks and Montanez 2006)). Understanding and harnessing the collective intelligence inherent in the mass collaboration of the Social Web is a challenging and important problem.

In this paper, we study the problem of uncovering latent structure in large-scale annotations. In particular, we propose a novel probabilistic generative model that views the aggregate social annotations applied to an object by a collaborative wide-scale distributed community of taggers as the product of a single underlying collective intelligence. By viewing the aggregate annotations as a community-based annotation document, the generative model can identify latent community-based "categories" of interest. These underlying categories of interest can be used to understand how tags are generated, to group semantically-related tags, to identify clusters of related documents, and so on.

As a case study, we apply the categorical annotation model to two prominent social tagging services – Flickr and Delicious – where we identify semantically-meaningful categories of interest. We further explore Delicious to understand the relationship between the annotations applied to a document and the content intrinsic to the document. We find that the proposed model identifies semantically coherent hidden categories that are complementary to the topics discovered through the application of a traditional content-based topic model. Based on this result, we illustrate an approach for integrating the annotation-based categorical model with content-based approaches for Web object exploration.

Background and Related Work

Social annotations have received growing research attention in the past few years. In this section, we provide a brief overview of some related work on (i) modeling and analyzing social annotations; and on (ii) text-based topic modeling, which inspires the annotation model introduced in this paper.

Analyzing social annotations

In one of the earliest studies of social tagging, Golder and Huberman (2005) found a number of clear structural patterns in Delicious, including the stabilization of tags over time, even in the presence of large and heterogeneous user communities. This stabilization (which might be counter-intuitive, especially in contrast to the tightly controlled metadata produced by domain experts) suggests a shared knowledge in tagging communities. These results are

echoed by Halpin et al. (2007), who found a power-law distribution for Delicious tags applied to Web pages – meaning that in the aggregate, distinct users independently described a page using a common tagging vocabulary. Similar results can be found elsewhere, including (Cattuto, Loreto, and Pietronero 2006), (Cattuto et al. 2007), (Li, Guo, and Zhao 2008), and (Veres 2006). Other work on tagging and incentives include (Sen et al. 2006) and (Marlow et al. 2006). These results motivate our interest in uncovering hidden categories that could help explain these phenomena.

Topic modeling

The annotation model presented in this paper is inspired by related work in text-based topic modeling. A topic model typically views the words in a text document as belonging to hidden (or "latent") conceptual topics. Prominent examples of latent topic models include Latent Semantic Analysis (LSA) (Deerwester et al. 1990), Probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999), and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). Topic models are an important component of many information retrieval and language modeling applications. There have been a number of extensions to traditional topic models including applications to hypertext (Gruber, Rosen-Zvi, and Weiss 2008) and email networks (McCallum, Corrada-Emmanuel, and Wang 2005).

Recently, there have been some efforts to adapt topic models to social annotations, including (Plangrasopchok and Lerman 2007; Wu, Zhang, and Yu 2006; Zhou et al. 2008). For example, in (Wu, Zhang, and Yu 2006), the authors propose a model to derive emergent semantics of tags, users, and content from a single underlying conceptual space. Similarly, in (Zhou et al. 2008), the authors propose an annotation model to unify a document's content with the tags applied to the document in the context of information retrieval. Our work differs from these previous efforts in at least two aspects. First, these models are tied to the text representation of the annotated document, and so cannot be easily extended to non-textual objects like images and videos. In contrast, we clearly distinguish the generation process that models an object's annotations from the generation process that models the object itself, so our model can be adapted to non-textual images and videos. Second, we model the annotation process as a collective decision that aggregates the behavior of many users, so the community-wide consensus dictates the mapping from resources to latent variables.

The Community-based Categorical Annotation (CCA) Model

In this section we propose a probabilistic generative model that aims to model the social annotation process. By modeling the communities of interest that engage in social tagging and the implicit categories that each community considers, we develop the Community-based Categorical Annotation (CCA) Model. The CCA model views a *category* as a mixture of tags and a *community* as a mixture of categories. Hence, a community of interest is inherently composed of the tags that it uses.

Reference model

We consider a universe of discourse $\mathcal{U}$ consisting of D socially annotated objects: $\mathcal{U} = \{O_1, O_2, ..., O_D\}$. We view each socially annotated object O_i by both its intrinsic content C_i and the social annotations S_i attached to it by the community of users. Hence, each object is a tuple $O_i = \langle C_i, S_i \rangle$ where the content and the social annotations are modeled separately. We call the social annotations S_i applied to an object its *social annotation document*. For example, the object corresponding to a Web page annotated in the Delicious community would consist of the HTML contents of the Web page as well as the *social annotation document* generated by the members of the Delicious community. A social annotation document can be modeled by the set of tags and their frequencies:[1] $S_i = \{\langle tag_j, freq(tag_j) \rangle\}$ In contrast to traditional Web pages and text documents that are typically written by a single author or a team working together, a social annotation document is "written" by contributors that are largely unaware of each other and the tagging decisions made by others. Questions remain – How are these social annotation documents produced? And what does this process tell us about the collective intelligence underlying these documents, and how can this knowledge impact information discovery and sharing?

Generating social annotations with CCA

We begin with an example. Suppose we have an image of a Tyrannosaurus rex. The collaborative tagging environment allows this object to be tagged by users with various interests, expertise, and in various human languages. Hence, the social annotation document associated with this image may include tags that were applied by a scientist e.g., tags like `cretaceous` and `theropod`), by an elementary school student (e.g., tags like `meat-eater` and `t-rex`) and by a French-speaking tagger (e.g., tags like `carnivore` and `lézard-tyran`).

We view the underlying groups that form around these interests, expertise, and languages as distinct *communities*. For each community, there may be some number of underlying *categories* that inform how each community views the world. Continuing our example, the scientist community may have underlying categories centered around Astronomy, Biology, Paleontology, and so on. For each object, the community selects tags from the appropriate underlying category or mixture of categories (e.g., for tagging the dinosaur, the tags may be drawn from both Biology and Paleontology).

In practice, these communities and categories are *hidden* from us; all we may observe is the social annotation document that is a result of these communities and the categories they have selected. Inspired by recent work on LDA and other text-based topic models (recall the Related Work

[1] As future work, it would be reasonable to additionally model the time at which each tag was applied, as well as the particular user applying the tag. Note that the CCA model introduced in the following section implicitly models users via tag co-occurrence.

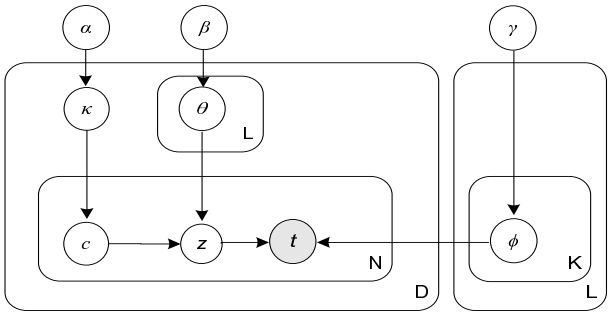

Figure 1: Graphical representation of the CCA Model.

section), we propose to model the generation of tags in the social annotation document using a generative probabilistic model called the Community-based Categorical Annotation (CCA) Model.

Formally, CCA assumes a corpus of D social annotation documents drawn from a vocabulary V of tags, where each social annotation document S_i is of variable length N_i. The model assumes that the tags in a social annotation document are generated from a mixture of L distinct communities, where each community is a mixture of hidden categories K_l, and where each category is a mixture of tags. Therefore, the tagging process involves two steps: 1) the selection of a community from which to draw tags and 2) the selection of the categories that influence the preference over tags based on the object's content, and the tagger's perception/understanding of the content. The CCA tag generation process is illustrated in Figure 1 and described here:

1. for each community $c = 1, ..., L$
   - for each category $z = 1, ..., K_c$
     - select V_c dimensional $\phi_z \sim$ Dirichlet(γ)
2. for each object S_i, $i = 1, ..., D$
   - Select L dimensional $\kappa \sim$ Dirichlet(α)
   - for each community $c = 1, ..., L$
     - select K_c dimensional $\theta_c \sim$ Dirichlet(β)
   - For each tag position $S_{i,j}$, $j = 1, ..., N_i$
     - Select a community $c_{i,j} \sim$ multinomial(κ_i)
     - Select a category $z_{i,j} \sim$ multinomial$(\theta_{c_{i,j}})$
     - Select a tag $t_{i,j} \sim$ multinomial$(\phi_{z_{i,j}}^{c_{i,j}})$

A social annotation document's community distribution $\kappa_i = \{\kappa_{i,j}\}_{j=1}^{L}$ is sampled from a Dirichlet distribution with parameter $\alpha = \{\alpha_i\}_{i=1}^{L}$. A community's category distribution $\theta_i = \{\theta_{i,j}\}_{j=1}^{K}$ is sampled from a Dirichlet distribution with parameter $\beta = \{\beta_i\}_{i=1}^{K}$. A category's tag distribution $\phi_z = \{\phi_{z,i}\}_{i=1}^{|V|}$ is sampled from a Dirichlet distribution with parameter $\gamma = \{\gamma_i\}_{i=1}^{|V|}$. The generative process creates a social annotation document by sampling for each tag position $S_{i,j}$ a community $c_{i,j}$ from a multinomial distribution with parameter κ_i, a category $z_{i,j}$ from a multinomial distribution with parameter $\theta_{c_{i,j}}$. A tag is then sampled for that position from a multinomial distribution with parameter $\phi_{z_{i,j}}^{c_{i,j}}$.

Based on the model, we can write the likelihood that a tag position $S_{i,j}$ in a social annotation document is assigned a specific tag t as:

$$p(S_{i,j} = t|\kappa_i, \Theta, \Phi) = \sum_{l=1}^{L}\sum_{k=1}^{K_l} p(S_{i,j} = t|\phi_k^l)p(z_{i,j} = k|\theta_l)p(c_{i,j} = l|\kappa_i).$$

Furthermore, the likelihood of the complete social annotation document S_i is the joint distribution of all its variables (observed and hidden):

$$p(S_i, z_i, c_i, \kappa_i, \Theta, \Phi|\alpha, \beta, \gamma) = \prod_{j=1}^{N_i} p(S_{i,j}|\phi_{z_{i,j}}^{c_{i,j}})p(z_{i,j}|\theta_{c_{i,j}})p(c_{i,j}|\kappa_i).$$

Integrating out the distributions κ_i, Θ, and Φ and summing over c_i and z_i gives the marginal distribution of S_i given the priors:

$$p(S_i|\alpha, \beta, \gamma) = \iiint p(\kappa_i|\alpha)p(\Theta|\beta)p(\Phi|\gamma) \times \prod_{j=1}^{N} p(S_{i,j}|\kappa_i, \Theta, \Phi)d\Phi d\Theta d\kappa_i$$

Finally our universe of discourse $\mathcal{U}$ consisting of all D social annotation documents occurs with likelihood:

$$p(\mathcal{U}|\alpha, \beta, \gamma) = \prod_{i=1}^{D} p(S_i|\alpha, \beta, \gamma)$$

Parameter estimation and inference

The CCA model provides a generative approach for describing how social annotation documents are constructed. But our challenge is to work in the reverse direction – taking a set of social annotation documents and inferring the underlying model (including the hidden community and category distributions). This entails learning model parameters κ, Θ, and Φ (the distributions over communities, categories, and tags, respectively).

Although exact computation of these parameters is intractable, several approximation methods have been proposed in the literature for solving similar parameter estimation problems (like in LDA), including expectation maximization (Blei, Ng, and Jordan 2003), expectation propagation (Minka and Lafferty 2003), and Gibbs sampling (Heinrich 2004). In this paper, we adopt Gibbs Sampling (Heinrich 2004) which is a special case of Markov-chain Monte Carlo methods that estimates a posterior distribution of a high-dimensional probability distribution. The sampler draws from a joint distribution $p(x_1, x_2, ..., x_n)$ assuming the conditionals $p(x_i|x_{-i})$ are known, where $x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$.

For community assignment c, category assignment z, tag assignment t of tag positions in a corpus, and given the parameters α, β and γ, Gibbs sampling computes:

$$p(c_i, z_i|c_{-i}, z_{-i}, t) \propto \frac{n_S^{c_i} - 1 + \alpha_{c_i}}{n_S - 1 + \sum_c \alpha_c} \times \frac{n_{z_i}^{t_i} - 1 + \gamma_{t_i}}{n_{z_i} - 1 + \sum_t \gamma_t} \times \frac{n_S^{z_i} - 1 + \beta_{z_i}}{n_S^{c_i} - 1 + \sum_z \beta_z}$$

where t_i is the tag at position i, z_i is the category, c_i is the community, S is the object, $n_S^{c_i}$ is the count of positions in the object assigned to community c_i, n_S is the length of the object, $n_{z_i}^{t_i}$ is the count of positions with category z_i and tag t_i in the corpus, n_{z_i} is the count of positions with category z_i in the corpus, and $n_S^{z_i}$ is the count of positions with category z_i in the object.

The first factor represents the weight of community c_i in the object , the second represents the contribution of the tag at position i to category z_i in the entire corpus, while the third factor represents the weight of category z_i in the object.

Having estimated the community assignment c and category assignment z, estimates of Φ,Θ and κ are computed as follows:

$$\phi_{z,t} = \frac{n_z^t + \gamma_t}{n_z + \sum_t \gamma_t}$$

$$\theta_{i,z} = \frac{n_S^z + \beta_z}{n_S^c + \sum_z \beta_z}$$

$$\kappa_{i,c} = \frac{n_S^c + \alpha_c}{n_S + \sum_c \alpha_c}$$

Now for a new unseen social annotation document $\tilde{S}$, the Gibbs sampler can predict its tag assignment as follows:

$$\phi_{z,\tilde{t}} = \frac{n_z^{\tilde{t}} + n_z^t + \gamma_t}{n_z + \sum_t \gamma_t}$$

where $n_z^{\tilde{t}}$ is the count of positions with category z and tag t in the unseen object. Its category distribution is:

$$\theta_{\tilde{S},z} = \frac{n_{\tilde{S}}^z + \beta_z}{n_{\tilde{S}}^c + \sum_z \beta_z}$$

where $n_{\tilde{S}}^z$ is the count of positions with category z in the unseen object, and its community distribution is:

$$\kappa_{\tilde{S},c} = \frac{n_{\tilde{S}}^c + \alpha_c}{n_{\tilde{S}} + \sum_c \alpha_c}$$

where $n_{\tilde{S}}^c$ is the count of positions with community c in the unseen object.

Applying CCA to Flickr and Delicious

Given the categorical annotation model, we next apply the model to two prominent social tagging services – Flickr (for images) and Delicious (for Web pages).

Flickr dataset: For Flickr, we began a crawl from the tag cloud at `http://flickr.com/photos/tags`. We have identified 1,578,437 images that have been annotated by 42,156 unique users who have used 156,127 unique tags. For the experiments in this paper, we considered a sample of 92,000 images that have been tagged by 44,980 unique tags. We normalize the data and train the categorical annotation model with 90,000 objects and use the rest for testing.

Delicious dataset: Like Flickr, the Delicious crawler starts with a set of popular tags. Our crawler has discovered 607,904 unique tags, 266,585 unique Web pages annotated by Delicious, and 1,068,198 unique users. Of the 266,585 total Web pages, we have retrieved the full HTML for 47,852 pages. We filter this set to keep only pages in English with a minimum length of 20 words, leaving us with 27,572 Web pages with 16,216 unique annotations. Since many of the pages annotated by Delicious are primarily text documents, we also parsed the text of each document for an analysis discussed later in the paper. We use $20,000$ of the objects to train our model and the remaining $7,572$ are used for testing.

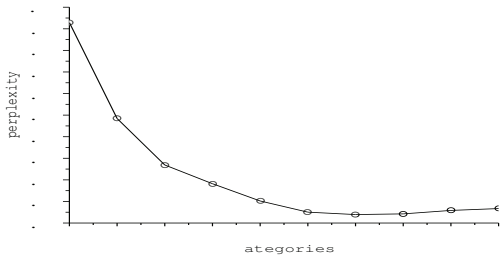

Figure 2: CCA-based category perplexity for Flickr.

Revealing hidden categories

One challenge to discovering latent structure in social annotations is to identify the appropriate number of hidden categories and hidden communities of interest that generated the observed data. Since the hidden categories and communities are not directly observed, we must use some unsupervised method.

In this section, we begin by considering the simplified case of a single community, but with an unknown number of hidden categories. We revisit this assumption in the following section. To identify the number of categories, we rely on a standard measure from information theory – perplexity. Perplexity measures how well a model (here the categorical annotation model built over a training set) predicts a test sample, and it is has been widely used in text-based topic modeling (e.g., (Blei, Ng, and Jordan 2003; Zhou et al. 2008)). We measure perplexity on a held-out set $\tilde{D}$ using the parameters of an estimated model $\mathcal{M}$ for a given dimension (or category) K for the hidden variable:

$$Perp(\tilde{D}) = \exp - \frac{\sum_{d=1}^{\tilde{D}} \log P(S_d|\mathcal{M})}{\sum_{d=1}^{\tilde{D}} N_d}$$

where

$$\log P(S_d|\mathcal{M}) = \sum_{t=1}^{V} n_d^{(t)} \log \left(\sum_{k=1}^{K} \phi_{k,t} \theta_{d,k} \right)$$

and $n_d^{(t)}$ is the number of times terms t was observed in document S_d and N_d is the length of S_d. The variable ϕ is a model parameter while the variable θ is computed for the held-out set. Low perplexity values indicate a good selection of the number of categories for the hidden variable given a corpus.

We experimented with different category dimensions for both Flickr and Delicious. The perplexity as a function of the number of categories for Flickr is shown in Figure 2.

Table 1: Flickr: 10 of the 70 discovered categories and the most likely tags per category (in order of $\phi_{z,t}$).

Cat 0:	boat, sport, itali, water, torino, athlet, ship, turin, sundai, sail, oar, rower, competit, ...
Cat 1:	canada, veteran, vancouv, memori, war, remembrancedai, dai, ontario, remembr, ...
Cat 2:	portrait, face, hand, woman, photoshop, hair, girl, color, lip, photograph, self, retrato, ...
Cat 3:	build, citi, architectur, old, urban, tower, histor, skyscrap, skylin, stone, center, librari, ...
Cat 4:	water, river, blue, reflect, bridg, fish, sky, boat, canon, artist, washington, mountain, ...
Cat 5:	mountain, winter, snow, landscap, lake, switzerland, cold, montagna, alp, trek, ...
Cat 6:	art, graffiti, paint, urban, streetart, street, tag, draw, sticker, illustr, abstract, artist, ...
Cat 7:	cat, anim, love, kitten, cute, kitti, pet, gato, felin, chat, gatto, bunni, rabbit, heart, ...
Cat 8:	train, railwai, tourist, tourism, station, laura, railroad, unitedkingdom, ride, york, locomot,...
Cat 9:	food, cook, cake, restaur, chocol, dinner, sweet, eat, minnesota, yummi, wine, bake, ...

Table 2: Delicious: 10 of the 40 discovered categories and the most likely tags per category (in order of $\phi_{z,t}$).

Cat 0:	webdesign, design, inspir, web, resource, templat, galleri, award, web2.0, websit, ...
Cat 1:	secur, financ, monei, .net, storag, invest, backup, asp.net, c#, busi, econom, bank, ...
Cat 2:	googl, mobil, calendar, phone, sync, api, voip, cellphon, comparison, nokia, sm, ...
Cat 3:	mac, osx, appl, wiki, softwar, ipod, macosx, app, applic, tool, ssh, wikipedia, quicksilv, ...
Cat 4:	educ, math, learn, resourc, teach, kid, technolog, mathemat, school, interact, elearn, ...
Cat 5:	tutori, howto, photoshop, tip, refer, guid, adob, articl, resourc, effect, trick, text, ...
Cat 6:	photographi, photo, imag, galleri, flickr, camera, slideshow, mindmap, stock, space, ...
Cat 7:	rubi, rail, rubyonrail, host, nyc, amazon, web, http, authent, s3, webhost, develop, ...
Cat 8:	fun, humor, funni, comic, cool, geek, interest, entertain, humour, del.icio.us, cartoon, ...
Cat 9:	video, visual, anim, movi, tv, film, youtub, motiongraph, motion, stream, media, ...

The horizontal axes show the number of categories and the vertical axes show the perplexity values. Notice the decrease in perplexity as the number of categories increase, as well as the different rates of decrease. For Delicious, we observe a similar curve, but with a "knee" at 40 categories. Based on these results, we selected 70 categories for Flickr and 40 categories for Delicious. Given the choice of the number of categories for both Flickr and Delicious, what are the discovered categories? And are they semantically coherent? In Table 1 and 2, we report the most significant annotations for a sample of 10 of the discovered categories in each dataset ranked by probability of tag given a category $\phi_{z,t}$. We find that overall the discovered categories appear to be semantically meaningful. As future work, it will be interesting to evaluate these discovered categories in a concrete application setting (e.g., tag-based information retrieval).

To further illustrate the revealed categories, we report in Table 3 the most relevant documents per category for 10 of the Delicious categories. We rank the documents using the probability of a category given a document $\theta_{i,z}$. We find that the quality of these results is consistent across categories.

Discovering communities

Given the results of uncovering hidden categories, we next turn to the nature of results when we estimate both the communities of interest (which recall are composed of underlying categories) and the categories within each community (which recall are composed of tags). Experimentally, we have run the CCA model with several community/category combinations, and in Table 4 we report a representative result for 5 communities and 5 categories for Delicious.[2]

[2]As part of our continuing work, we are developing techniques to optimize both the number of categories and communities. Due to the space constraint, we omit that discussion here.

Note how the communities of interest are centered around categories that share some thematic relationship. For example, Comm2 is a "Lifestyle" community of interest with categories related to shopping, travel, food, and books. In the flat single community analysis of the previous section, these types of categories would either be combined into a single category of interest, blurring the distinct interests of each category, or the categories may be separated but not linked by community. Here, we see how the CCA model provides a hierarchical layer for grouping related categories by their common community of interest. Further, note that the two more technically minded communities are indeed quite distinct – Comm3 is centered around "Web 2.0" from a consumer point-of-view (with categories related to YouTube, blogs, and social networking), whereas Comm4 is centered around "Web 2.0" from a development point-of-view (with categories related to different web development tools and languages). These results are encouraging and in our continuing work, we are exploring techniques to further refine the quality of community formation.

Categories vs. Content-Based Topics

Now that we have seen how the CCA model can identify hidden categories and communities of interest that are used to drive the social annotation process, we revisit the relationship between an object's content and its social annotation document (recall $O_i = \langle C_i, S_i \rangle$). Previous efforts have unified these two views to generate both the content and the annotations through a single process (e.g., (Wu, Zhang, and Yu 2006; Zhou et al. 2008)). The intuition is that the author of a document and the social annotators of a document are driven by the same motivations. Indeed, there is evidence that many tags applied to a Web page can also be found in the text of the page (Heymann, Koutrika, and Garcia-Molina 2008).

Table 3: Top 4 Most Relevant Documents per Category ranked by $\theta_{i,z}$ (showing 10 of the 40 categories)

Category 0 (Web design)	**Category 5 (Photoshop)**
http://www.webbyawards.com/webbys/current.php?season=12	http://psdtuts.com/photo-effects-tutorials/applying-a-realistic-tattoo/
http://www.coolhomepages.com/	http://abduzeedo.com/creating-smoke
http://vandelaydesign.com/blog/galleries/minimal-websites-designs/	http://psdtuts.com/text-effects-tutorials/create-a-spectacular.../
http://www.designlicks.com/flash/index.php	http://psdtuts.com/tutorials-effects/seriously-cool-photoshop.../
Category 1 (Banking and money)	**Category 6 (Photography)**
https://www.fidelity.com/	http://hirise.lpl.arizona.edu/earthmoon.php
http://home.ingdirect.com/	http://www.boston.com/bigpicture/2008/05/cassini_nears_four.../
http://www.chase.com/	http://wildphoto.smugmug.com/
http://www.wamu.com/personal/default.asp	http://www.boston.com/bigpicture/2008/06/martian_skies.html
Category 2 (Calendar syncing and messaging)	**Category 7 (Ruby)**
http://www.gcalsync.com/	http://ec2onrails.rubyforge.org/
http://oggsync.com/	http://code.macournoyer.com/thin/
http://www.clickatell.com/pricing/message_cost.php	http://www.hostingrails.com/
http://www.daveswebsite.com/software/gsync/	http://mongrel.rubyforge.org/
Category 3 (Apple/Mac)	**Category 8 (Fun and humor)**
http://www.magnetk.com/expandrive	http://www.dilbert.com/
http://macntfs-3g.blogspot.com/	http://www.achewood.com/
http://code.google.com/p/macfuse/	http://xkcd.com/162/
http://www.sccs.swarthmore.edu/users/08/mgorbach/MacFusionWeb/	http://www.sarcasmsociety.com/
Category 4 (Education)	**Category 9 (Video and movies)**
http://school.discoveryeducation.com/schrockguide/assess.html	http://www.netflix.com/MemberHome
http://www.learningpage.com/	http://www.netflix.com/
http://edhelper.com/	http://joox.net/
http://www.teach-nology.com/	http://www3.alluc.org/alluc/

Table 4: Communities, their categories, and the most likely tags per category (in order of $\phi_{z,t}$).

Comm 0	Cat 0	art, design, paper, drawing, diy, fun, cool, animation, toys, crafts...
	Cat 1	humor, funny, fun, comics, geek, blog, comic, humour, cool, webcomic...
	Cat 2	dictionary, reference, language, english, writing, tools, thesaurus, slang ...
	Cat 3	games, game, fun, flash, gaming, free, online, puzzle, secondlife, charity...
	Cat 4	imported, misc, firefoxbookmarks, bookmarks, firefoxtoolbar, myspace, computer...
Comm 1	Cat 0	science, math, kids, mathematics, reference, education, astronomy, games, physics,...
	Cat 1	howto, productivity, lifehacks, tips, gtd, reference, diy, tutorial, blog, organization...
	Cat 2	education, learning, resources, teaching, elearning, technology, free, video, language, online...
	Cat 3	photography, photo, photos, images, flickr, art, tools, graphics, free, image...
	Cat 4	web20, tools, wiki, collaboration, presentation, mindmap, online, software, powerpoint, free...
Comm 2	Cat 0	shopping, environment, design, green, tshirts, home, sustainability, clothing, activism, shop...
	Cat 1	travel, reference, maps, airline, world, guide, flights, seating, airlines, geography...
	Cat 2	books, reference, library, research, history, literature, ebooks, free, archive, writing...
	Cat 3	news, blog, technology, magazine, politics, blogs, tech, culture, daily, media...
	Cat 4	food, cooking, recipes, blog, recipe, music, reviews, reference, blogs, howto...
Comm 3	Cat 0	tools, web20, mobile, productivity, software, collaboration, calendar, phone, widgets, web...
	Cat 1	video, tv, streaming, videos, movies, media, youtube, free, television, online...
	Cat 2	business, marketing, advertising, startup, internet, technology, trends, entrepreneurship, ideas, media...
	Cat 3	web20, social, community, socialnetworking, twitter, collaboration, tools, socialsoftware, networking, web...
	Cat 4	blog, web20, blogs, blogging, news, rss, aggregator, web, tools, technology...
Comm 4	Cat 0	javascript, ajax, programming, framework, development, web20, web, library, webdev, yahoo...
	Cat 1	wordpress, blog, themes, theme, plugin, blogging, plugins, blogs, templates, design...
	Cat 2	php, opensource, cms, software, web, email, development, drupal, ecommerce, programming...
	Cat 3	css, webdesign, web, design, html, reference, tutorial, webdev, standards, development...
	Cat 4	programming, java, development, reference, c, net, database, tutorial, sql, tools...

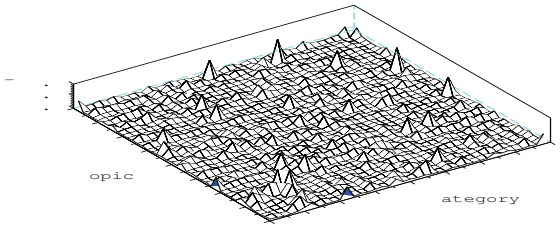

Figure 3: Topic vs. Category Similarity

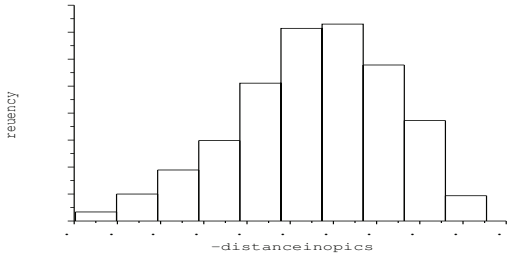

Figure 4: Jensen-shannon distance distribution in categories: Objects with < 0.1 JS-distance in Category space

Such a unified view, however, would seem to be meaningful for annotated objects that are primarily text (like Web pages). It is less clear how to unify the content and annotation generation process for non-textual objects like images and videos. Hence, we next study whether the unified document content + social annotation model is even reasonable for primarily text-based Web pages.

Categories and topics on Delicious

For the Delicious dataset, we considered the 40 categories discovered using the CCA model. We additionally ran LDA (Blei, Ng, and Jordan 2003) on the document content of the collected Web pages and identified 40 latent topics (again using perplexity). We are interested to understand if the underlying topic modeling approach for generating a document is the same as the categorical modeling approach for generating a social annotation document.

To measure the similarity of the content and annotation generation processes, we compare all pairs of topics and categories. If the two processes are similar, we would expect to see many similar topic/category pairs. For each possible pair of categories and topics, we measured their similarity using the Jensen-Shannon distance (Lin 1991) for comparing two probability distributions p and q over an event space X: $JS(p,q) = 0.5\,[KL(p,m) + KL(q,m)]$ where $m = 0.5(p+q)$ and $KL(p,q)$ is the Kullback-leibler divergence defined as: $KL(p,q) = \sum_{x \in X} p(x) \cdot \log(p(x)/q(x))$.

To compute the JS-distance between a a topic and category we represent each topic or category z by a probability vector ϕ_z over the union of the tag vocabulary space and the content vocabulary space. In Figure 3 we compare all (topic,category) pairs. The x-axis shows the categories, the y-axis shows the topics, and the z-axis shows (1−JS-distance). We use (1−JS-distance) for visibility where similar pairs will show as large spikes on the plot.

While there are some clear spikes, for the majority of topics there is no clear mapping to related categories, and vice versa. Hence we believe that the categorical annotation model identifies semantically coherent hidden categories that are not the same as the topics discovered through the application of a traditional content-based topic model – which further validates the need to separately model and study the collective intelligence annotation process from the content-generation process.

To further understand this separation, we also examined the set of social annotation document pairs that are categorically similar, where we considered pairs with JS-divergence less than 0.1 in the categorical space. How topically similar are these documents? Do documents that share similar tags also share similar content? In Figure 4, we report the JS-divergence between these categorically-similar objects over their *content-based topic similarity*. Note how many of these categorically-similar Web pages are quite dissimilar in topic space. In other words, objects tagged with similar tags do not necessarily have similar content.

Conversely, we also considered the set of Web page pairs in our Delicious dataset that had a JS-divergence less than 0.1, where we measured the JS-divergence over the topics associated with each document. We find that many of these topically-similar Web pages are quite dissimilar in categorical space. These results echo what we saw in Figure 4, that two documents may share many keywords in common (i.e., are topically similar), but their view from the community of social annotations is quite different.

Browsing in topic and category space

Finally, we briefly illustrate one way to use both the annotation-based categorical model and the content-based topic approach for discovery and exploration of Web objects. The main idea is to explore objects based both on their categorical and topical similarity (and dissimilarity) to a candidate query object. Here, we consider an example Web page in the Delicious dataset concerned with the popular 1980s-era Rubik's Cube and several methods for solving the puzzle. The vocabulary for this page is overwhelmingly mathematical and based solely on the content this document is classified under the mathematics topic with high probability. However, the document also clearly belongs to the games and puzzles category (and this is reflected in the tags assigned to it). Given this query document, in Figure 5 we show the most relevant documents to our query document based on three views: (i) similar in topic space and similar in category space – these documents are primarily mathematical approaches to Rubik's cube and similar puzzles; (ii) similar in topic space, but dissimilar in category space – these documents are primarily mathematical documents; and (iii) dissimilar in topic space, but similar in category space – these documents are primarily about games and puzzles.

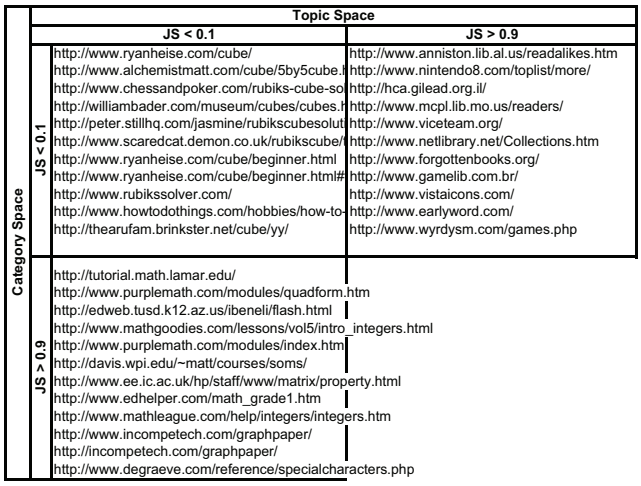

Figure 5: Browsing in Category and Topic spaces

Conclusions

Understanding and modeling the collective semantics centered around large-scale social annotations is a promising research avenue with potential implications for information discovery and knowledge sharing. As a step in this direction, we have presented a new community-based categorical model for generating social annotations. Based on this model, we showed how to discover latent structure in large-scale social annotations collected from Delicious and Flickr. In our continuing work, we are considering more fine-grained hierarchical models of the social annotation process and extending the integrated browsing model, as well as the scope of our experimental validation to other social tagging communities.

Acknowledgments

We would like to thank Robert Graham for helping with the Delicious crawl. This work is partially supported by faculty startup funds from Texas A &M University and the Texas Engineering Experiment Station.

References

Bao, S.; Xue, G.; Wu, X.; Yu, Y.; Fei, B.; and Su, Z. 2007. Optimizing web search using social annotations. In *WWW '07*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, 993–1022.

Brooks, C. H., and Montanez, N. 2006. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW'06*.

Cattuto, C.; Baldassarri, A.; Servedio, V. D. P.; and Loreto, V. 2007. Vocabulary growth in collaborative tagging systems.

Cattuto, C.; Loreto, V.; and Pietronero, L. 2006. Collaborative tagging and semiotic dynamics.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W.; and Harshman, R. A. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6):391–407.

Golder, S., and Huberman, B. A. 2005. The structure of collaborative tagging systems.

Gruber, A.; Rosen-Zvi, M.; and Weiss, Y. 2008. Latent topic models for hypertext. In *UAI*, 230–239. AUAI Press.

Halpin, H.; Robu, V.; and Shepherd, H. 2007. The complex dynamics of collaborative tagging. In *WWW '07*.

Heinrich, G. 2004. Parameter estimation for text analysis. Technical report.

Heymann, P.; Koutrika, G.; and Garcia-Molina, H. 2008. Can social bookmarking improve web search? In *WSDM '08*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR '99*, 50–57.

Li, R.; Bao, S.; Yu, Y.; Fei, B.; and Su, Z. 2007. Towards effective browsing of large scale social annotations. In *WWW '07*.

Li, X.; Guo, L.; and Zhao, Y. E. 2008. Tag-based social interest discovery. In *WWW '08*, 675–684.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Trans. on Info. Theory* 37(1).

Marlow, C.; Naaman, M.; Boyd, D.; and Davis, M. 2006. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HT '06*.

McCallum, A.; Corrada-Emmanuel, A.; and Wang, X. 2005. The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Workshop on Link Analysis, Counterterrorism and Security*, 33–44.

Minka, T., and Lafferty, J. 2003. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 352–359.

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

Sen, S.; Lam, S. K.; Rashid, A. M.; Cosley, D.; Frankowski, D.; Osterhouse, J.; Harper, M. F.; and Riedl, J. 2006. Tagging, communities, vocabulary, evolution. In *International Conference on Computer Supported Cooperative Work (CSCW)*.

Veres, C. 2006. The language of folksonomies: What tags reveal about user classification. *Natural Language Processing and Information Systems* 58–69.

Wu, X.; Zhang, L.; and Yu, Y. 2006. Exploring social annotations for the semantic web. In *WWW '06*, 417–426.

Zhou, D.; Bian, J.; Zheng, S.; Zha, H.; and Giles, C. L. 2008. Exploring social annotations for information retrieval. In *WWW '08*, 715–724.