The Topic-Perspective Model for Social Tagging Systems

Caimei lu, Xiaohua Hu, Xin Chen, Jung-ran Park
College of Information Science and Technology
Drexel University, Philadelphia, PA, USA
caimei.lu@drexel.edu
{xiaohua.hu, xin.chen, jung-ran.park}@ischool.drexel.edu

ABSTRACT
In this paper, we propose a new probabilistic generative model, called Topic-Perspective Model, for simulating the generation process of social annotations. Different from other generative models, in our model, the tag generation process is separated from the content term generation process. While content terms are only generated from resource topics, social tags are generated by resource topics and user perspectives together. The proposed probabilistic model can produce more useful information than any other models proposed before. The parameters learned from this model include: (1) the topical distribution of each document, (2) the perspective distribution of each user, (3) the word distribution of each topic, (4) the tag distribution of each topic, (5) the tag distribution of each user perspective, (6) and the probabilistic of each tag being generated from resource topics or user perspectives. Experimental results show that the proposed model has better generalization performance or tag prediction ability than other two models proposed in previous research.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *Parameter learning*; H.1.2 [**Models and Principles**]: User/Machine Systems – *Human factors, Human information processing*

General Terms
Algorithms, Experimentation, Human Factors, Design.

Keywords
Social tagging, Social annotation, User modeling, Perplexity.

1. INTRODUCTION
The evolution from Web 1.0 to Web 2.0 is accompanied by the popularity of various Web-based user services and applications. The rapidly growing social data created by users through these Web 2.0 services and applications has intrigued active research. Lots of research work has been done focusing on how to utilize the social data to improve the traditional data mining and information retrieval (IR) tasks. Social annotations are one type of such social data. They are free-form tags generated by users on social tagging systems or social bookmarking applications. A recent study reports that around 115 million social bookmarks were available in 2008 on the social bookmarking website *dell.icio.us* [9]. As a new type of information source, social annotations have been exploited in recent literature for various application purposes, such as tag recommendation or prediction [10, 11, 25], clustering [16, 21], classification [27], and information retrieval [28].

In these studies, the value of social annotation data is generally exploited from two aspects. Firstly, social annotations themselves are viewed as additional metadata about the described resources, and used to enrich document representation in data mining and IR tasks. Secondly, the social tagging network or the structured relationships among users, social annotations and resources are more extensively exploited for purposes like classification, clustering and IR. Compared to the metadata property of social annotations, the social network formed through users' social tagging behavior provides richer and more valuable information for learning the topics of web resources, the semantics of tags and user communities.

In order to utilize the network information in social tagging systems, it is prerequisite to properly model the relationships among the different entities involved in the social tagging systems, including users, resources, tags, and content units of resources. Several methods have been proposed to model the social tagging network. For instance, in [14], social tagging system is described as a tripartite network with users, tags and annotated items as three types of nodes. The tripartite network of social tagging was projected into bipartite and unipartite networks in order to discover its underlying structures. In [17] social tagging system is also modeled as a tripartite graph which extends the traditional bipartite model of ontologies with a social dimension (users). [23] investigates the network characteristics of social tagging system using metrics adapted from classical network measures, including characteristic path length and clustering coefficient . [27] proposes a social tagging graph in which tags act as bridges to connect resources from heterogeneous domains. A semi-supervised classification algorithm is then developed based on the social tagging graph. All these methods represent social tagging systems in a flat network structure.

Another approach to model the interactions among different objects in social tagging systems is to simulate the generation process of social annotations with a generative model. For instance, Wu et al. propose a probabilistic generative model in which the three entities in a social tagging system (tags, resources, and users) are mapped to a common conceptual space, which is represented by a multi-dimensional vector with each dimension

KDD'10, July 25–28, 2010, Washington DC, USA.

corresponding to a knowledge category [26]. Besides, hierarchical Bayesian models based on Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) have also been proposed to model the social tagging process [13, 20, 21].

In this paper, we propose a new probabilistic generative model based on the well-known LDA model for simulating the generation process of social annotations. Different from other models, in our model, the tag generation process is modeled separately from the content term generation process. This design is based on the observation that tags are generated differently from document content terms: the content terms contained in a resource are generated by a single or a small group of authors sharing common interests, while the social tags of a resource are generated by many users with different interests, expertise and purposes. A user's perspective not only always differs from the authors' but may also distinct from another user's perspective. The generation of a tag is always influenced by the user's perspectives. A set of tags applied to a resource can reflect both users' perspectives and the resource's topics.

In our model, we use two different latent variables to represent resource topics and user perspectives. Note that the perspective of a user not merely refers to the user's interest, but also covers the user's expertise, motivation, language and other personal factors. First, resource topics are generated and represented with word distributions. Then the identified topics of a resource are used to generate the tags together with user perspectives. Accordingly, we call our model "Topic-Perspective (TP) Model". By estimating this model, we can get more informative outputs than any other models proposed before. These outputs include: (1) the topical distribution of each document, (2) the perspective distribution of each user, (3) the word distribution of each topic, (4) the tag distribution of each topic, (5) the tag distribution of each user perspective, (6) and the probabilistic of each tag being generated from resource topics or user perspectives.

A distinct feature of our model is that, during the tag generation process, resource topics and user perspectives together generate the social tags for a resource, and each tag differs in the extent of depending on resource topics or user perspectives. The rationality of this design is evidenced by the existence of social tags with various functional purposes. Golder and Huberman identify seven different functions of tags (shown in the first column in Table 1) [7]. Based on Golder and Huberman's schema, Bischoff et al. specify eight categories of tags (shown in the second column of Table 1) [3]. Sen et al. summarize the seven tag functions proposed by Golder and Huberman into three categories: Factual tags, Subjective tags, and Personal tags [24]. Intuitively, Factual tags or the tags of the first five categories identified in [3] are more closely related to resource content and extrinsic to the taggers, while the Subjective tags and Personal tags are less connected to resource topics and more influenced by users' perspectives. For instance, on *del.icio.us*, the Factual tags for the URL (http://www.brainyquote.com/) titled "Famous Quote and Quotations at BrainyQuote" include tags like *quotes*, *quotations*, *writing*, *literature*, etc. Meanwhile it is also annotated with the Subjective and Personal tags such as *funny*, *cool*, *interesting*, etc. Factual tags are more valuable than Subjective and Personal tags for representing resource content, and thus are more effective when used as additional information in data mining and information retrieval tasks. Therefore, it is necessary to identify this property of a tag, namely whether it relies more on resource topics or user factors. Based on the proposed generative model for social annotation, we are able to estimate the probabilistic that each tag is generated from user perspectives or resource topics.

The rest of the paper is organized as follows: Section 2 reviews related work on topic modeling and various generative models for social annotations proposed in previous research. Section 3 presents the proposed TP Model for social annotation and introduces the parameter estimation process. In section 4, based on a social booking dataset crawled from *del.icio.us*, we evaluate the performance of the proposed model and compared it with other two models described in previous research. Section 5 discusses future work. We conclude the paper in section 6.

Table 1. The mapping of three different classification schemas of social tags

Golder et al. [7]	Bischoff et al.[3]	Sen et al.[24]
What or who it is about	Topic	Factual
Refining categories	Time	
	Location	
What it is	Type	
Who owns it	Author/Owner	
Qualities and characteristics	Opinions/Qualities	Subjective
Task organization	Usage context	Personal
Self reference	Self reference	

2. RELATED WORK

2.1 Topic Analysis using Generative Models

In the data mining and information retrieval community, a set of effective probabilistic models have been proposed to simulate the generation of a document, such as the Naïve Bayesian model, the Probabilistic Latent Semantic Indexing (PLSI) model [11] and the Latent Dirichlet Allocation (LDA) model [2]. Particularly, the LDA model has become popular in the text mining community due to its solid theoretical foundation and promising performance. Since it was first proposed, a wide variety of its extensions have been proposed in different contexts for different modeling purposes.

Many extended LDA models have explored information other than document words for topic learning. For instance, the author-topic model proposed in [22] uses the authorship information together with the words to learn topics. The correlated LDA model learns topics simultaneously from images and caption words [1, 5]. The switchLDA model reveals topics from content words and entities in news articles [19], the Link-LDA model and Topic-Link LDA model represents topics and author communities using both content words and links between documents [6, 15], etc. Most of these models can also be applied in the social annotation context when considering the tags as the additional information source for topic learning.

2.2 Generative Models for Social Tagging

A variety of LDA-based generative models have been proposed for modeling the generation of social annotations. For clustering purpose, Ramage et al. propose a LDA model which jointly models the generation of content word and tags [21]. This model is essentially the same as the Conditionally-independent LDA (CI-LDA) model used for generating words and entities in [19] and the Link LDA model used for generating words and document links in [6]. Figure 1(a) shows the graphical representation of this model. We can see that in CI-LDA model, the tag is generated

from the same source as the word: the topic of the document. Users' impact on the generation of tags is not considered in this model. This is not appropriate, because the process that generates content is different from the annotation process, especially for non-textual resources like images and videos. A more appropriate generative model should consider the role of users in tag generation process.

Zhou et al. propose a more comprehensive model for social annotation, which considers the impact of both document topic and user interest on tag generation [28]. In this model, a word in a document is generated in the same way as in the standard LDA model. On the annotation side, each tag is generated similarly. First, a user decides to annotate a web document and then the user selects a topic, based on which a tag is generated to describe the document. Although this model provides a comprehensive view about the generation process of both content words and tags, unfortunately, it involves many parameters to be estimated, which make the model hardly tractable. Therefore, the authors of [28] propose a simplified LDA annotation model for social annotation by assuming that words and tags are both generated from the same topics shared by documents and users. Obviously, this is not proper. As mentioned, document words are created by the authors of the document, while tags are generated by a large group of users with different background and interests.

Kashoob et al. propose a generative model called community-based categorical annotation (CCA) model [13]. Different from other models, in this model, the annotation process is modeled as a collective decision of user communities, which are viewed as groups forming around interests, expertise, language, etc. It is assumed that each community has a number of underlying categories as its world view. Each category can be represented as a mixture of tags. Therefore, in CCA model, a tag in a document is generated from a category which is further generated from a community selected for the document. The outputs of this model include the community distribution of a resource, category distribution of a community, and tag distribution of a category. The authors compare the category distribution generated from the CCA model and the hidden topics in content words generated separately through standard LDA model, and conclude that tag-based categories are not the same as content-based topics. A defect of this model is that it ignores the dependence of tags on resource topics. According to this model, tags are generated independently from resource topics. Apparently, this is not the case in real tag generation process. Moreover, this model only considers the collective impact of communities on social annotation. However, in some cases like personalized search, we are more interested in the information about individual users.

A recent work adapts the correlated or correspondence LDA (CorrLDA) model proposed in [4] for social annotation. The model is graphically represented in Figure 1(b). The CorrLDA model first generates word topics for a document. Then the topics associated with the words in the document are used to generate tags. Compared to the CI-LDA model, the CorrLDA model can force a greater degree of correspondence between two information sources (in this case, words and tags). But like CI-LDA, the user information is missed in the tag generation process. In order to incorporate user factors into the tag generation process, another model called User-Topic-Tag Model is proposed in [4]. In this model, users are treated like authors in the author-topic model proposed in [22]. First, for each word in the document, a user is chosen uniformly at random from the group of users who annotated the documents. Then, a topic is chosen from a distribution over topics specific to that user, and the word is generated from the chosen topic. Finally, as in the CorrLDA model, the topics associated with the words in the document are used to generate each tag associated with the document. Although this model accounts for user factors, it does not correctly simulate the real social annotation process because users are modeled as creators of content words instead of tags.

The Topic-Perspective model proposed in this paper overcomes the limitations of previous models by simulating the real social annotation process and representing all related entities (users, documents, words, and tags) and latent variables (topics, user perspectives) in a unified model.

3. TOPIC-PERSPECTIVE MODEL

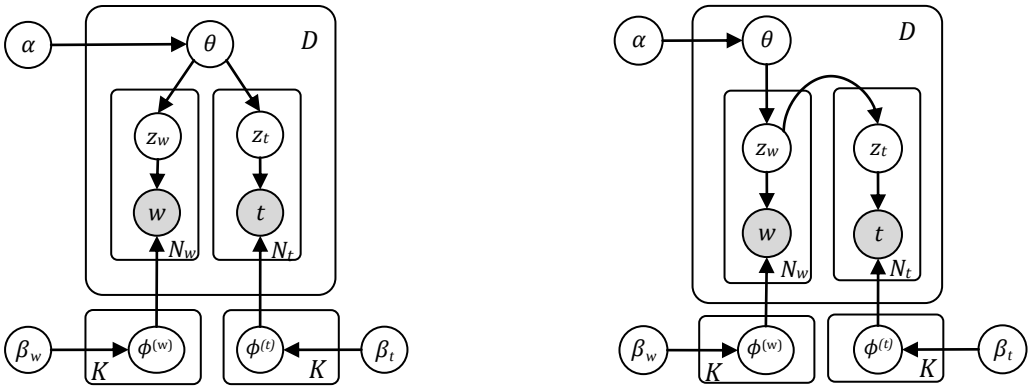
In this section, we introduce the proposed Topic-Perspective Model for social annotation. This model depicts the social annotation process and the generation process of content terms in a unified framework. The motivation behind this model is to represent and connect all the observed and hidden variables in a unified framework. By estimating this model, we can learn the topical structure of the documents, terms, and tags, the tagging perspectives of users and the representation user perspectives with tags at the same time. The output of this model can be used to enhance text mining and IR performance. For instance, the identified user perspectives can be utilized for personalized search. We can incorporate the user perspectives into the search process on both the query side and the document side. On the query side, user perspective can help us decide the exact information need described by the query term proposed by the user. On the document side, the perspective distributions of users who have ever annotated a document can help measure the relevance of the annotated document to the query.

3.1 Model Formulation

The model is designed based on the real social annotation process. Before a tag was created by a user for a document, the document terms already exist. Therefore, we first consider the term generation process which can be modeled with a standard LDA model. Then when a user annotates a document, two factors act on the tag generation process. One is the topics of the document; the other is the perspective adopted by the user. Even for the tags created by the same user, the extent that each tag is affected by user's perspectives may be different, because, as mentioned, each tag can be created for different functional purposes. For instance, some tags are created to specify the topics of the resources, while other tags may be created for self-reference, quality evaluation, and opinion expression. Intuitively, the tags of the former kind are more dependent on the topics of the documents, while the latter kind are more affected by users' personal perspectives.

In order to reflect this nature of social annotations in the generative model, we adopt a switch variable to control the influence of user perspectives and document topics on tag generation. The proposed model is illustrated in Figure 2. The meanings of notations used in Figure 2 are summarized in Table 2.

As shown in Figure 2, this model primarily comprises of two parts split by the dash line. The right part is essentially the standard LDA, which models the generation of content terms contained in the documents. For each word w in a document d, a topic z is first sampled, and then the word w is drawn conditioned on the topic.



(a) CI-LDA (b) CorrLDA

Figure 1. Graphical representation of (a) CI-LDA and (b) CorrLDA for modeling social annotations

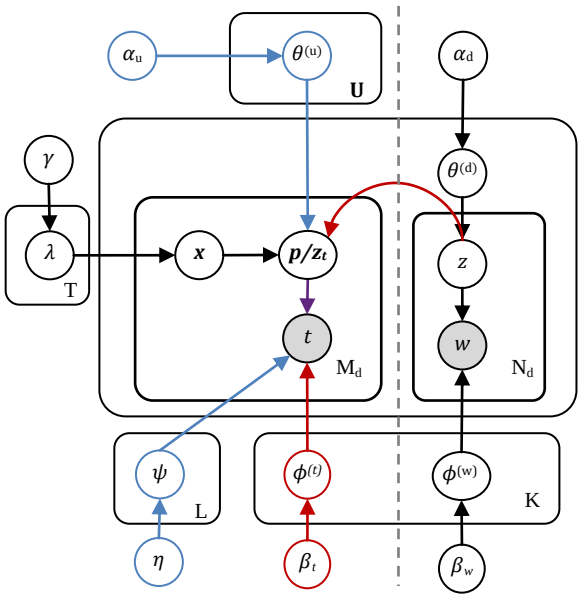


Figure 2. The Topic-Perspective Model for Social Annotation

The document d is generated by repeating the process N_d times, which is the number of word tokens in d. The left part of Figure 2 models the generation of tags. Each tag t created by user u for document d can be drawn from either the topics associated with d's content words or u's perspectives. To decide the source of each tag, a switch variable x is introduced. The value of x (which is 0 or 1) is sampled based on a binomial distribution λ (with a Beta prior γ). When the sampled value of x equals 1, tag t is drawn from the topic z_t which is uniformly sampled from the topics learned from the words in document d. The red arrows in Figure 2 show this process. When x equals 0, a perspective p is first sampled from the perspective distribution (θ_u) for user u, and then the tag t is drawn from the tag distribution ψ_p of perspective

p. The blue arrows in Figure 2 illustrate this procedure. Overall, the generation process of words and tags in the Topic-Perspective model can be described as follows:

1) For each of the D documents d, sample $\theta^{(d)}{}_d \sim$ Dirichlet (α_d);
2) For each of the U users u, sample $\theta^{(u)}{}_u \sim$ Dirichlet(α_u);
3) For each of the K topics k, sample $\phi^{(w)}{}_k \sim$ Dirichlet(β_w), and sample $\phi^{(t)}{}_k \sim$ Dirichlet(β_t);
4) For each of the L user perspectives l, sample $\psi_l \sim$ Dirichlet(η);
5) For each of the N_d word tokens w_i in document d:
   a) sample a topic $z_i \sim$ Multinomial ($\theta^{(d)}{}_d$);
   b) sample a word $w_i \sim$ Multinomial ($\phi^{(w)}{}_{z_i}$);
6) For each of the T tags t in the collection D, sample $\lambda_t \sim$ Beta (γ);
7) For each of the M_d tag tokens t_j in document d created by user u;
   a) sample a flag $X \sim$ Binomial (λ_{tj});
   b) if (X = 1):
      i) Sample a topic $z_{tj} \sim$ Uniform $(z_{w_1}, \cdots, z_{w_{N_d}})$;
      ii) Sample a tag $t_j \sim$ Multinomial ($\phi^{(t)}{}_{zj}$);
   c) if (X = 0):
      i) Sample a perspective $p_j \sim$ Multinomial(θ_u);
      ii) Sample a tag $t_j \sim$ Multinomial (ψ_{pj});

3.2 Parameter Estimation

The Topic-Perspective has six parameters for estimation: (1) the document-topic distribution $\theta^{(d)}$, (2) the topic-word distribution $\phi^{(w)}$, (3) the topic-tag distribution $\phi^{(t)}$, (4) the user-perspective distribution $\theta^{(u)}$, (5) the perspective-tag distribution ψ, (6) and the binomial distribution λ. Several methods have been developed for estimating the latent parameters in LDA model, such as the variational expectation maximization [2], expectation propagation [18], and Gibbs sampling [8, 19]. Compared to the other two methods which are very computationally expensive, Gibbs sampling often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA. Therefore we select this approach for parameter estimation. In the Gibbs Sampling algorithm for the standard LDA model, a Markov chain is constructed and converges to the posterior distribution on topic z. The transition between successive states in the Markov chain is modeled by repeatedly drawing a topic for each observed word from its conditional probability. For our model, during the Gibbs Sampling procedure an additional Markov chain is introduced for simulating the tag generation. Inspired by the Gibbs Sampling equation for standard LDA model [28], we derive

the sampling equations for our model. The major notations used in the following equations are explained in Table 2.

- Sampling equation of the word topic variables for each content word w_i. (The same as standard LDA model) :

$$P(z_i = k \mid w_i = v, z_{-i}, w_{-i}, \alpha_d, \beta_w) \propto \frac{C^{KD}_{kd,-i} + \alpha_d}{\sum_{k'} C^{KD}_{k'd,-i} + K\alpha_d} \bullet \frac{C^{WK}_{vk,-i} + \beta_w}{\sum_{v'} C^{WK}_{v'k,-i} + V\beta_w} \quad (1)$$

- Sampling equation of the tag topic variables when the switch variable X =1:

$$p(x_j = 1, z_j^{(t)} = \tilde{z} \mid t_j = q, z_{-j}, t_{-j}, \beta_w, \beta_t, \gamma) \propto \frac{\tilde{n}_{q,-j} + \gamma}{n_q + \tilde{n}_{q,-j} + 2\gamma} \bullet \frac{C^{KD}_{\tilde{z}d}}{N_{w_d}} \bullet \frac{C^{TK}_{q\tilde{z},-j} + \beta_t}{\sum_{q'} C^{TK}_{q'\tilde{z},-j} + T\beta_t} \quad (2)$$

- Sampling equation of the tag perspective variables when the switch variable X =0:

$$p(x_j = 0, p_j = l \mid t_j = q, p_{-j}, t_{-j}, \alpha_u, \beta_t, \gamma) \propto \frac{n_{q,-j} + \gamma}{n_q + \tilde{n}_{q,-j} + 2\gamma} \bullet \frac{C^{LU}_{lu,-j} + \alpha_u}{\sum_{l'} C^{LU}_{l'u,-j} + L\alpha_u} \bullet \frac{C^{TL}_{ql,-j} + \beta_t}{\sum_{q'} C^{TL}_{q'l,-j} + T\beta_t} \quad (3)$$

After a set of sampling processes based on the posterior distributions calculated with the above equations, we can estimate the parameters for any single sample using the following equations:

$$\theta^{(d)}_{kd} = \frac{C^{KD}_{kd,-i} + \alpha_d}{\sum_{k'} C^{KD}_{k'd,-i} + K\alpha_d}, \quad \phi^{(w)}_{vk} = \frac{C^{WZ}_{vk} + \beta_w}{\sum_{v'} C^{WZ}_{v'k} + V\beta_w}$$

$$\phi^{(t)}_{q\tilde{z}} = \frac{C^{TK}_{q\tilde{z},-j} + \beta_t}{\sum_{q'} C^{TK}_{q'\tilde{z},-j} + T\beta_t}, \quad \theta^{(u)}_{lu} = \frac{C^{LU}_{lu,-j} + \alpha_u}{\sum_{l'} C^{LU}_{l'u,-j} + L\alpha_u}$$

$$\psi_{ql} = \frac{C^{TL}_{ql,-j} + \beta_t}{\sum_{q'} C^{TL}_{q'l,-j} + T\beta_t}, \quad \lambda_q = \frac{\tilde{n}_{q,-j} + \gamma}{n_q + \tilde{n}_{q,-j} + 2\gamma}$$

4. EXPERIMENTS AND RESULTS

In this section, we investigate the performance of the proposed Topic-Perspective LDA (TP-LDA) model based on a social bookmarking dataset crawled from *del.icio.us*. We also compare our model with two other LDA-based generative models for social annotation: the CI-LDA model and CorrLDA model mentioned in section 2. We choose these two models for comparison, because like our model, they do the topical analysis of words and tags simultaneously. Actually, our model is built on the CorrLDA. It extends the CorrLDA by incorporating the user factors in the tag generation process. The generation process of these two models is graphically represented in Figure 1. For details about the Gibbs sampling process and equations of these two models, readers can refer to [19], where they are used for modeling the topics of words and entities in news articles.

4.1 Datasets

The dataset used for experiment is from a social tagging dataset we collected from the *del.icio.us* website during January 2009 and February 2009. The original dataset contains 3,246,424 posts for 1,731,780 URLs created by 4784 users. We selected 45,462 URLs which are indexed in the human-maintained Web directory ODP (Open Directory Project). Then we crawled the web content of these URLs. After filtering out web pages with no tags or containing less than 20 words, 41190 webpage documents remained. To further clean the dataset, stopwords and words with term frequency less than 5 are filtered out. Besides, phrase tags are also preprocessed. Because *del.icio.us* does not allow space within a tag, tags containing more than one word (phrase tags) are formed in a variety of ways. For instance, "java programming" may be formulated as "java_programming", "javaprogramming", "java-programming" or "java.programming, etc. In our experiment, we treat the phrase tags composed by the same terms but in different ways as the same tag. The final dataset used for experimentation contains 41190 documents, 4414 users, 28740 unique tags, and 129908 unique words. Then we randomly selected 10% of the documents and their associated users and tags as a held-out test data and trained the model on the remaining 90%.

Table 2. Notations

d, u, v, q, k, l	the instance of a variable: *d* for document, *u* for user, *v* for word, *q* for tag, *k* for topic, *l* for perspective
D, U, W, T	total number of documents, users, words, and tags in the dataset.
K, L	The selected number of topics and perspectives.
N_d, M_d	The number of word tokens and tag tokens contained in document d
$C^{KD}_{kd,-i}$	the number of times that topic k has occurred in document d, except the current instance
$C^{WK}_{vk,-i}$	the number of times word *v* is assigned to topic *k*, without counting the current instance.
$C^{TK}_{q\tilde{z},-j}$	the number of times tag *q* is generated from topic *k*, without counting the current instance.
$C^{LU}_{lu,-j}$	the number of times that perspective *l* is adopted by user *u*, except the current instance.
$C^{TL}_{ql,-j}$	the number of times tag *q* is generated from perspective *l*, without counting the current instance.
$\tilde{n}_{q,-j}$	the number of times that tag *q* is generated from topics (x_q=1), except current assignment;
$n_{q,-j}$	the number of times that tag *q* is generated from perspectives (x_q=0), except current assignment;
$\theta^{(d)}$	a $D \times K$ matrix indicating document-topic distribution.
$\phi^{(w)}$	a $K \times W$ matrix indicating topic-word distribution
$\phi^{(t)}$	a $K \times T$ matrix indicating topic-tag distribution
$\theta^{(u)}$	a $U \times L$ matrix indicating user- perspective distribution
ψ	a $L \times T$ matrix indicating perspective-tag distribution
λ	a vector indicating the probability that each tag is generated from topics.
α_d, α_u, β_w, β_t, η, γ	hyperparameters and priors of Dirichlet distributions.

4.2 Evaluation Criterion

In out experiment, we use perplexity as the criterion for model evaluation. Perplexity is a standard measure for evaluating the generalization performance of a probabilistic model. The value of perplexity reflects the ability of a model to generalize to unseen data. Specifically, in our case, perplexity reflects the ability of a model to predict tags for new unseen documents. The perplexity is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean of per-word (per-tag in our case) likelihood [2]. A lower perplexity score indicates better generalization performance. Formally, the perplexity for a test set of D_{test} documents is calculated as follows:

$$perplexity(D_{test}) = \exp\{\frac{\sum_{d=1}^{D_{test}} \log(p(t_d))}{\sum_{d=1}^{D_{test}} M_d}\}, \quad (4)$$

$$p(t_d) = p(x_{t_d} = 1)\sum_{k=1}^{K} p(t_d \mid z_k)\, p_{test}(z_k \mid d) + p(x_{t_d} = 0)\sum_{l=1}^{L} p(t_d \mid p_l)\, p_{test}(p_l \mid u)$$

In the above equation t_d is a tag included in the test document d. The probabilities $p(t_d \mid z_k)$, $p(t_d \mid p_l)$,and $p(x)$ are learned from the training process, and $p_{\text{test}}(z_k \mid d)$ and $p_{\text{test}}(p_l \mid u)$ are estimated through a Gibbs Sampling process on the test data based on the parameters $\phi^{(w)}$, $\phi^{(t)}$, ψ, and λ learned from training data.

4.3 Experimental Setup

The Topic-Perspective model has six Dirichlet prior parameters. We test several values for each parameter and found that their effect on the perplexity value is little. So we set α_d=0.3, α_u =0.3, β_w=0.05, β_t =0.05, η=0.05, γ=0.5 for all experiments.

The remaining question is how to select the number of topics K and the number of perspectives L. We first fix the number of perspectives to a certain number, and then test the perplexity of the trained model on the test data for different topic numbers. The smallest topic number which leads to the minimum or near minimum perplexity is selected. After the topic is chosen, the perspective number is selected similarly based on the perplexity. Figure 3 shows a plot of perplexities on five different settings of K, when the perspective number is fixed to 80. We can see that in general the perplexity scores for all topic number settings decrease along the iterations. The algorithm tends to converge after about 40 iterations. Along the iterations, larger setting of topic number always leads to smaller perplexity value from the start, and indicating a better prediction performance. But the effect of increase in topic number on perplexity value gets smaller when the topic number gets larger. When the topic number set to 160, the perplexity value actually goes up. Therefore, we set the topic number K=80 which leads to the minimum perplexity among the five settings. The situation for selecting perspective number is similar. Figure 4 displays the plot of perplexities for five settings of perspective number L when topic number is set to 80. Still the perspective number L=80 leads to the minimum perplexity score. And when L increases to 160, the perplexity value sharply goes up. So for the final experiment, both topic number and perspective number are set to 80.

4.4 Results

4.4.1 Tag Perplexity

We compare the tag prediction abilities of our Topic-Perspective model with CorrLDA model and CI-LDA model based on the perplexity value. Figure 5 plots the perplexity results for each model over different topic numbers. The perspective number of TP-LDA is set to 80, and the iteration numbers for all three models are set to 80. We can see that, before K=80 TP-LDA constantly performs better than other two models especially when the topic number is small. For all three models, larger topic number generally leads to smaller perplexity scores. This is because the increased topic number reduces the uncertainty in training. However, the effect of topic number on the three models' performance is different. TP-LDA model is lest affected by the topic number. Especially, when the topic number increases to 160, its perplexity value grows up. This is because TP-LDA incorporates the users' perspective information into the tag generation process, and the predicted tags do not completely accounts on document topics.

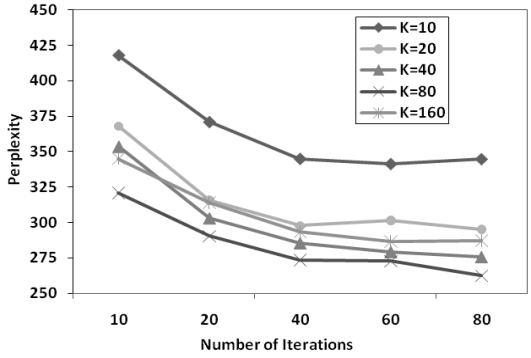


Figure 3. The perplexities over the iterations for five settings of topic number when perspective number L=80

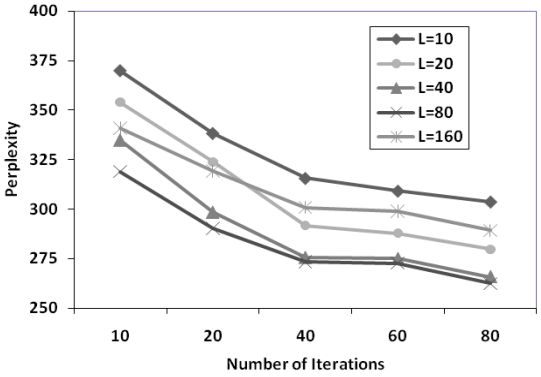


Figure 4. The perplexities over the iterations for five settings of perspective number when topic number K=80

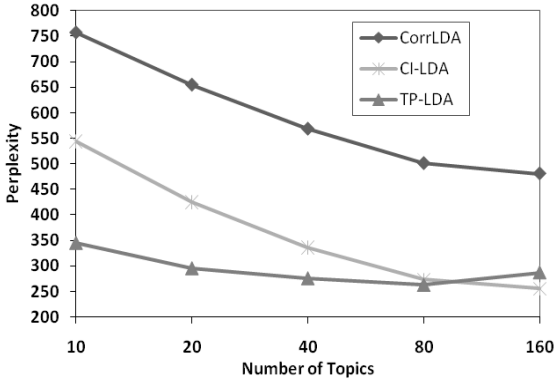


Figure 5. The perplexity results of CorrLDA, CI-LDA and TP-LDA (Topic-Perspective Model) for topic number K=10, 20, 40, 80, 160

From figure 5, we can also see that CorrLDA performs worse than CI-LDA. This is because CI-LDA uses both content words and tags to learn document topics, but CorrLDA only learns topics from content words. Recall that, in CorrLDA, only the topics learned from content words are used to generate tags. Therefore, in CI-LDA model, the topics learned from the training data are more associated with tags and thus are more effective for tag prediction. This result further indicates the difference of words and tags in topical structure. Although CI-LDA generates better

results, experimental results show that CI-LDA's word topics and tag topics are too decoupled. Little correspondence can be found between the words and tags generated to represent the same topic. The TP-model overcomes this limitation without sacrificing the performance.

4.4.2 Discovered topics and perspectives

Because no quantitative measures are available, we evaluated the topics and perspectives discovered by our model by examining the top words and tags assigned to each topic and the top tags assigned to each perspective.

Despite of the lack of quantitative assertions, we observe generally high semantic correlations among the top words and tags for the each topic, and high correspondence between the words and tags for the same topic. The themes of the discovered topics are diverse, and mostly related to the hot subjects, such as web design, programming, traveling, shopping, education, politics, etc. Table 3 displays the top ten words and tags of a random subset of discovered topics. Because of the coherent semantics of the words and tags for each topic, the theme of each topic is obvious. For instance, Topic 7 is about the war and politics, Topic 13 is on outdoor activities, Topic 15 is associated with movies, and so on.

Our model also discovers the user perspectives from the tags. The perspectives are more complicated than topics. The correlation among the tags assigned to each perspective is not as obvious as those for topics. This is because, unlike topic, a user perspective does not reflect a pure aspect of tags. Each perspective may combines several user factors of social tagging, such as user' domain background, preference, interest, motivation, etc. Despite of the complexity of perspectives, we can still identify some patterns by examining the tags assigned to each perspective. Table 4 lists a subset of discovered perspectives and their top tags. We can see that the tags assigned to perspectives are very different from those assigned to topics. If we look back on Bischoff's classification of tags in Table 2, it is apparent that the tags assigned to perspectives generally belong to the categories other than Topic. For instance the tags for Perspective 11 are mostly for describing the documents' type, tags in Perspective 24 are used for opinion-expression and self-reference, and tags for Perspective 64 are used for task organization and self-reference.

4.4.3 The generation sources of Tags

In our model, we use an additional variable λ ($0< \lambda <1$) to record the probability that each tag is generated from topics or user perspectives. Greater value of λ indicates a higher probability that the tag is generated from document topics and vice versa. Table 5 lists some example tags with λ = 1, 0.5 and 0. The tags with λ=1 are completely generated from topics and not affected by users' perspectives. We can see that these tags can clearly and objectively reflect the topics of the annotated documents. Contrarily, the tags with λ = 0 are totally generated by users' perspectives. It is clear to see that these tags contribute little to reflecting the topics of the annotated documents. They are created by users for other purposes other than identifying topics. An interesting observation was made on the tags with λ = 0.5, which are equally influenced by document topics and user perspectives. From table 5, we can see that these tags are mostly terms invented by the users which cannot be found in dictionaries. Most of them are phrases with no space between words, such as "audiomagazine". Different from tags with λ = 0, these tags are actually related to the document topics. In other words, they are created to describe the topics in a personal way.

Table 3. A subset of discovered topics

Topic ID	Top words	Top tags
7	war world militaries nation state force govern unite iraq countries international israel american armies peace	politics history world international war military activism poverty information africa islam government middleeast europe humanright
13	mountain fish camp boat adventure sea river park trail climb ski new lake gear sail	travel camp backpack hike photography climb sail photo knot boat gear nature adventure ski kayak
15	movie film star video dvd man episode new release trailer love review girl fan season	movy video entertainment film music review movie humor television medium fun funny cinema stream comicstrip
16	church god christian beer bible jesus religion faith new catholic christ life religion holi john	religion bible christianity christian church history buddhism mythology atheism theology spirituality philosophy apologetics catholic culture
20	window linux software file microsoft install mac computer user server program your run desktop disk	software window linux mac freeware osx utility ubuntu apple backup download sysadmin virtualization security opensource
25	law legal copyright inform public patent right state court govern lawyer act protect file agency	law copyright legal government internet privacy patent technology security politics research right p2p tech plagiarism
28	recipe food cook cup coffee chocolate cake eat tea cheese bake add bread make water	food recipe cook howto health drink coffee nutrition vegetarian collection restaurant tea bake kitchen diet
46	university student school education studies college science teacher program teach course institution graduate academ department	education teach learn school research science elearn university resource academic college kid study math lessonplan
49	book write author stories publish writer read fiction comic chapter novel poetries edit amazon poem	book write ebook literature comicstrip read publish library comic poetry tutorial webcomic selfpublish author scifi fiction

Table 4. A subset of discovered perspectives

Perspective ID	Top tags
11	reference guide multimedia list help codec portal emulator comparison boot upload anonymous virtual organize proxy
24	competition switch event blogroll likeddesign inspire wysiwyg creative artistresource ria tagthese domainname affiliate editorial cooky
32	conference metadata openacce tag association folksonomy censorship preservation digitallibrary sheetmusic librarian secondlife rfid directory digitalgame
36	search link portal directory list system indie tag rock about customize current label usenet ezine kaizen synchronization
47	bookmark readlate quickd engl401 ircbot meetup shirt favoritesmenu emergent nikon twincity tattoo punk oreilly jobsite simplicity
51	good publication product stuff thesis awesome mypublication gobacktothis florida giztag sidebar nicedesign travelinfo longdistancetrip sourdough myprofile
52	developer article example onlinekit issuetrack flstudio aggregation swiftteam webbuild backend asus wtf guideline communication swiftmobile
58	compute archive multimedia wireless app macintosh directory communication datum codec freedom admin cheat classic mystuff list
64	toread todo totry todownload webdevelope mind tobrowse tobuy tocheck conference landscape frequentlyuse epge usefulsoftware intelligence

Table 5. Example Tags with three different values of λ

λ=1	λ=0.5	λ=0
library shop	palmpre audiomagazine	tag app interest
internet research	mathsware postapocalyptic	archive toread
socialnetwork	educause vomit nwiqpartn	datum code todo
statistic ruby ajax	singlespe masterproef	webservice
javascript webdev	richmullin sundial selenium	directory list guide
culture music	showstep webmath	link portal training
health graphic math	randynewman immortalism	site track article
security firefox cs	malazan architecturalproduct	reference web20
politics recipe	fotologserevista biblioteque	online search tool
photography	caribbean europeana	free cool

5. FUTURE WORK

For future work, we want to apply the results of our model for tag prediction. Given a new document, based on the parameters estimated by our model, the tags can be predicted in two ways. For general tag prediction, we can only consider the tags with high topic probability and filter out tags with high perspective probability. The likelihood of a tag t for a test document d is:

$$p(t \mid d) = \sum_{k=1}^{K} p(t \mid z_k)\, p(z_k \mid d)$$

$p(t_d \mid z_k)$ is given by the topic-tag distribution, and $p(z_k \mid d)$ is estimated online based on the parameters learned from the training process. If we wants to predict the tags that could be created for a document by a specific user u who has also appeared in the training dataset, we can calculate the likelihood of a tag t for a test document d as follows:

$$p(t \mid d, u) = p(\lambda_t) \sum_{k=1}^{K} p(t \mid z_k)\, p(z_k \mid d) + [1 - p(\lambda)] \sum_{l=1}^{L} p(t \mid p_l)\, p(p_l \mid u)$$

,in which $p(t \mid z_k)$, $p(t \mid p_l)$, $p(p_l \mid u)$ and λ are learned from the training process, while $p(z_k \mid d)$ are estimated online.

We also plan to use the parameters learned from the model to enhance the performance of information retrieval (IR). For general information retrieval, we can smooth the IR language model with the tags of high topical probability and the topical structures of the words and tags. For personalized information retrieval we can further expand the IR language model with tags of high perspective probability and the users' perspectives.

6. CONCLUSIONS

In this paper, we propose a Topic-Perspective LDA model to simulate the tag generation process. By modeling the tag generation and word generation process separately and incorporating the user information into the tag generation process, the proposed model is able to model the social annotation system in a more meaningful way and achieve better generalization performance than other models. Besides, this model also generates useful information about the topical structures of tags and words, as well as the influences of document topics and user perspectives on different tags. The results derived from this model can be utilized for automatic tag recommendation, information retrieval and other text mining applications.

7. ACKNOWLEDGMENTS

This work is supported by NSF CCF 0905291 and NSFC 90920005 "Chinese Language Semantic Knowledge Acquisition and Semantic Computational Model Study".

8. REFERENCES

[1] D.M. Blei, and M.I. Jordan, Modeling annotated data The 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, Toronto, Canada, 2003, pp. 127-134.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022.

[3] K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu, Can All Tags be Used for Search?, CIKM'08, Napa Valley, California, USA, 2008, pp. 203-212.

[4] M. Bundschus, S. Yu, V. Tresp, A. Rettinger, M. Dejori, and H.-P. Kriegel, Hierarchical Bayesian Models for Collaborative Tagging Systems, ICDM '09. Ninth IEEE International Conference on Data Mining., IEEE, Miami, Florida, 2009, pp. 728-733.

[5] X. Chen, C. Lu, Y. An, and P. Achananuparp, Probabilistic models for topic learning from images and captions in online biomedical literatures, the 18th ACM conference on Information and knowledge management, ACM, Hong Kong, China 2009, pp. 495-504.

[6] Elena Erosheva, S. Fienberg, and J. Lafferty, Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences 101 (2004) 5220-5227.

[7] S. Golder, and B.A. Huberman, Usage Patterns of Collaborative Tagging Systems. Journal of Information Science 32 (2006) 198-208.

[8] T.L. Griffiths, and M. Steyvers, Finding scientific topics. Proceedings of National Academy of Sciences of the United States of America 101 (2004) 5228-5235.

[9] P. Heymann, G. Koutrika, and H. Garcia-Molina, Can Social Bookmarking Improve Web Search?, WSDM'08, Palo Alto, California, USA, 2008.

[10] P. Heymann, D. Ramage, and H. Garcia-Molina, Social Tag Prediction, SIGIR'08, Singapore, 2008, pp. 531-538.

[11] T. Hofmann, Probabilistic Latent Semantic Analysis, 15th Conference on Uncertainty in Artificial Intelligence, UAI'99 Morgan Kaufmann, Stockholm, Sweden, 1999.

[12] R. Jaschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme, Tag Recommendations in Folksonomies, Knowledge Discovery in Databases: PKDD 2007, 2007, pp. 506-514.

[13] S. Kashoob, J. Caverlee, and Y. Ding, A Categorical Model for Discovering Latent Structure in Social Annotations, The 3rd International AAAI Conference on Weblogs and Social Media San Jose, CA, 2009.

[14] R. Lambiotte, and M. Ausloos, Collaborative tagging as a tripartite network, arXiv:cs/0512090v2, 2005.

[15] Y. Liu, A. Niculescu-Mizil, and W. Gryc, Topic-link LDA: joint models of topic and author community, the 26th Annual International Conference on Machine Learning, ACM, Montreal, Quebec, Canada, 2009 pp. 665-672

[16] C. Lu, X. Chen, and E.K. Park, Exploit the Tripartite Network of Social Tagging for Web Clustering, CIKM'09, ACM, HongKong, China, 2009, pp. 1545-1548.

[17] P. Mika, Ontologies are us: A unified model of social networks and semantics. Journal of Web Semantics 5 (2007) 5-15.

[18] T. Minka, and J. Lafferty, Expectation-propagation for the generative aspect model, the 18th Conference in Uncertainty

in Artificial Intelligence, Morgan Kaufmann, Edmonton, Alberta, Canada, 2002, pp. 352-359.

[19] D. Newman, C. Chemudugunta, and P. Smyth, Statistical entity-topic models, the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Philadelphia, PA, 2006, pp. 680 - 686

[20] A. Plangprasopchok, and K. Lerman, Exploiting Social Annotation for Automatic Resource Discovery, AAAI-07 Workshop on Information Integration on the Web, arXiv.org, Vancouver, BC, Canada, 2007.

[21] D. Ramage, P. Heymann, C.D. Manning, and H. Garcia-Molina, Clustering the Tagged Web, WSDM 2009, ACM, Barcelona, Spain, 2009, pp. 54-63.

[22] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, The author-topic model for authors and documents, the 20th conference on Uncertainty in artificial intelligence AUAI Press, Banff, Canada 2004, pp. 487 - 494.

[23] C. Schmitz, M. Grahl, A. Hotho, G. Stumme, C. Cattuto, A. Baldassarri, V. Loreto, and V.D.P. Servedio, Network Properties of Folksonomies, WWW2007, ACM, Banff, Canada, 2007.

[24] S. Sen, S.K.T. Lam, A.M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F.M. Harper, and J. Riedl, Tagging, communities, vocabulary, evolution, CSCW'06, Banff, Alberta, Canada, 2006.

[25] Y. Song, L. Zhang, and C.L. Giles, A Sparse Gaussian Processes Classification Framework for Fast Tag Suggestions, CIKM'08, ACM, Napa Valley, California, USA, 2008, pp. 93-102.

[26] X. Wu, L. Zhang, and Y. Yu, Exploring Social Annotations for the Semantic Web, WWW 2006, ACM, Edinburgh, Scotland, 2006.

[27] Z. Yin, R. Li, Q. Mei, and J. Han, Exploring social tagging graph for web object classification, the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Paris, France, 2009, pp. 957-966

[28] D. Zhou, J. Bian, S. Zheng, H. Zha, and C.L. Giles, Exploring Social Annotations for Information Retrieval, WWW 2008, Beijing, China, 2008, pp. 715-724.