Exploring Collaboratively Annotated Data for Automatic Annotation

Jiafeng Guo, Xueqi Cheng, Huawei Shen, Shuo Bai
Institute of Computing Technology, CAS
Beijing, P.R.China
{guojiafeng,shenhuawei}@software.ict.ac.cn, cxq@ict.ac.cn, sbai@sse.com.cn

ABSTRACT

Social tagging, as a collaborative form of annotation process, has grown in popularity on the web due to its effectiveness in organizing and accessing web resources. This paper addresses the issue of automatic annotation, which aims to predict social tags for web resources automatically to help future navigation, filtering or search. We explore the collaboratively annotated data in social tagging services, in which collaborative annotations (i.e., combined social tags of many users) serve as a description of web resources from the crowds' point of views, and analyze its three important properties. Based on these properties, we propose an automatic annotation approach by leveraging a probabilistic topic model, which captures the relationship between resources and annotations. The topic model is an extension of conventional LDA (Latent Dirichlet Allocation), referred as Word-Tag LDA, which effectively reflects the generative process of the collaboratively annotated data and models the conditional distribution of annotations given resources. Experiments are carried out on a real-world annotation data set sampled from del.icio.us. Results demonstrate that our approach can significantly outperform the other baseline methods in automatic annotation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance, Theory

Keywords

social tagging, automatic annotation, topic model

1. INTRODUCTION

Annotating resources with descriptive keywords or tags is a common way to help organize and access large collections of resources. In the recent years, social tagging, as a collaborative form of the annotation process, is gaining popularity on the web. Social tagging services (e.g., Delicious[1] and Citeulike[2]) enable users to not only freely tag the web resources but also publicly share these information with others. Such collaborative annotations (i.e., combined social tags of many users) actually reflect how the crowds describe and organize the web resources. Therefore, a natural question is that whether we can learn from the "wisdom of the crowds", the collaboratively annotated data, to predict social tags for web resources automatically. This is what we addressed by automatic annotation here. Note that in this paper, we focus on document resources that are consisted of words, which represent a major type of resources on the web. Automatic annotation is potentially useful for wide applications including resource discovery, tag recommendation, web page search and clustering. For example, a large amount of un-annotated web resources can be automatically annotated to help discover similar resources to the existing annotated resources, which users might be interested in. Moreover, predicted tags might also be reasonable to suggest to users to assist their annotation process.

There is much previous work focused on social tagging, including analysis of tagging systems [9, 18], study of tag visualization [7], recommendation of tags [19, 14, 21], and discovery of hidden semantics in tags [22]. However, in this paper, we investigate exploiting collaboratively annotated data for automatic annotation. Such a problem fits in the framework of Semantic Web [2], which aims to automatically assign metadata to web resources. However, the primary approaches on Semantic Web either presume that an ontology is built before annotation [17, 12] or rely on text keywords extraction [5]. We propose to annotate resources with social tags, which are freely generated by web users without any a-priori formal ontology to conform to.

Our work is inspired by the following three important properties of the collaboratively annotated data in social tagging services: (1) **Stable**, which means that collaborative annotations for a particular web resource give rise to a stable distribution of social tags. (2) **Consistent**, which means that collaborative annotations reflect the content of web resources with its own literature. (3) **Shareable**, which means that most tags in collaborative annotations are shared by a large variety of web resources. Based on these properties, we propose an automatic annotation approach by leveraging a probabilistic topic model, which learns the relationship between resources and collaborative annotations. Specifically, the topic model is an extension of conventional LDA (Latent Dirichlet Allocation) [4], referred as Word-Tag LDA, which captures the conditional relationships between lower-dimensional representations (i.e. hidden topics) of words (in resources) and tags (in annotations). The approach simulta-

[1]http://del.icio.us
[2]http://www.citeulike.org

CAW2.0 2009, April 21, 2009, Madrid, Spain.
.

neously models the way how the web resource is generated as well as the way how the resource is subsequently annotated with collaborative annotations. With the Word-Tag LDA model learnt in hand, we are able to predict the conditional distribution of social tags given a new web resource. Note here we are interested in predicting a ranked list of social tags for web resources, not only a yes/no decision for each possible tag. Obviously, there are always a great amount of tags which can be assigned to web resources. Therefore, predicting a ranked list rather than a set of social tags makes more sense and is more effective in describing web resources.

We demonstrate the effectiveness of our automatic annotation approach by comparing with three types of baseline approaches, including keyword approach, similarity approach and joint approach. Experiments are conducted on a real-world annotation data set sampled from Delicious. Experimental results show that our model can significantly outperform all the other baselines in automatic annotation.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 introduces the problem of automatic annotation and conducts some preliminary studies. Section 4 describes the proposed Word-Tag LDA in details. Experimental results are presented in Section 5. Conclusions are made in the last section.

2. RELATED WORK

2.1 Research on Social Annotation

Social tagging has received major attention in the recent literature. The early discussion of the social tagging can be found in [9, 18, 20, 11]. In [9], Golder et al. provided empirical study of the tagging behavior and tag usage in Delicious. Quintarelli [18] gave a general introduction of social annotation and suggested that we should take it as an information organizing tool.

Research based on social tagging has been done in various areas such as emergent semantics [16, 22], tag recommendation [19, 14, 21], visualization [7], and information retrieval [1, 23]. Wu et al. [22] explored emergent semantics from the social annotations in a statistical way and applied the derived semantics to discover and search shared web bookmarks. Sigurbjörnsson et al. [19] proposed a method which uses tag co-occurrence to suggest additional tags that complement user-defined tags of photographs in Flickr. Heymann et al. [14] predicted tags for web pages based on the anchortext of incoming links, web page content, surrounding hosts and other tags applied to the URL. In [7], Dubinko et al. presented a novel approach to visualize the evolution of tags over the time. Bao et al. utilized social annotations to benefit web search [1]. They introduced SocialPageRank, to measure page authority based on its annotations, and SocialSimRank for similarities among annotations.

Different from the above work, we investigated the capability of social tagging in automatic annotation by utilizing the correlation between web resources and collaborative annotations.

2.2 Research on Automatic Annotation

Automatically assigning metadata to web resources is a key problem in Semantic Web area. A lot of work has been done about this topic. Early work [17, 12] mainly focused on the utilization of an ontology engineering tool. An ontology is usually built first and then manual annotation for web resources is conducted with the tool. In order to help automate the manual process, many approaches have been proposed and evaluated. Chirita et al. proposed P-TAG, a method which suggests personalized tags for web pages based on both the content of the page and the data on the user's desktop [5]. Dill et al. [6] presented a platform for large-scale text analysis and automatic semantic tagging. They proposed to learn from a small set of training examples and then automatically tag concept instances on the web. Handschuh et al. [13] considered the web pages that are generated from a database, and proposed to automatically produce sematic annotations from the database with an ontology. However, the primary approaches on Semantic Web either presume that an ontology is built before annotation or rely on text keywords extraction. We proposed to assign social tags to content, which are freely generated by users without any a-priori ontology to conform to.

3. PRELIMINARY STUDIES

The idea of a social approach to the semantic annotation is enlightened and enabled by the now widely popular social tagging services on the web. These social tagging services provide users a convenient collaborative form of annotation process, so that they can not only freely tag the web resources but also publicly share these information with others. Since social tagging has become an effective way of organizing and accessing large collections of web resources, a natural question is that whether we can predict social tags for web resources automatically to help future navigation, filtering or search. This is the automatic annotation problem addressed in our paper.

Before we discuss the technical aspects of our work, it is worth spending a moment to conduct some preliminary studies on social tagging. There are many popular social tagging services on the web, e.g. the social bookmarking systems [11]. Here we conduct some studies on a typical social bookmarking system, Delicious. In Delicious, an annotation activity (i.e. bookmark) typically consists of four elements: a user, a link to the web resource, one or more tags, and a tagging time. We define an annotation as

(User, Url, Tag, Time).

In this paper, we focus on the collaboratively annotated data, in which collaborative annotations (i.e., combined social tags of many users) serve as a description of web resources from the crowds' point of views. Therefore, we disregard the roles of User and Time, combine Tags of the same URL together, crawl the web resource (i.e. document words) referred by the URL and obtain the pair of (Words, Tags) as the collaboratively annotated data. For our preliminary studies, we randomly sampled 70,000 annotated URLs from Delicious and collected the corresponding word-tag pairs. We made some investigation over such data and explore the following three important properties, which become the foundation of our work:

(1) **Stable**, which means that collaborative annotations for a particular web resource give rise to a stable distribution of social tags. Although individual's annotation may have personal preference and varying tag vocabulary, collaborative annotations form a stable pattern in which the tag frequency distribution is nearly fixed. Fig. 1 shows this property, which was obtained by plotting the Kullback-Leibler (KL) divergence of tag frequency distribution for each step

Table 1: The top 10 frequent keywords and collaborative annotations of an example URL.

Url	http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html
Top *tf* Words	rss xml feed format schema web content feeds news share
Top Tags	rss xml syndication web reference programming web2.0 article tutorial howto

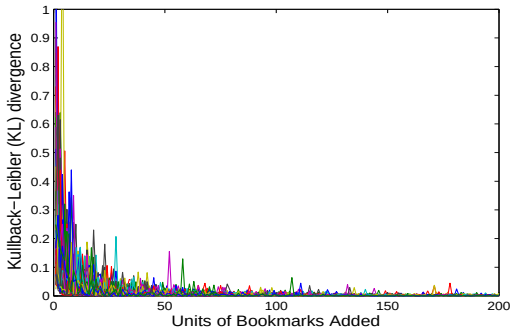

Figure 1: Kullback-Leibler (KL) divergence of tag frequency distribution at each step (as units of bookmarks added), with respect to the final tag distribution.

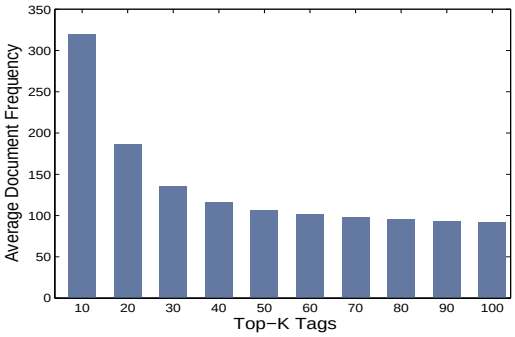

Figure 2: Average Document Frequency (DF) of top frequent tags.

(as units of bookmarks added) with respect to the final tag distribution for the resource (where "final" denotes distribution at the time the measurement was taken, or the last observation in the data). From Fig. 1 we can see that empirically after the first 100 or so bookmarks, the KL divergence between each step and the final one converges to zero (or close to zero), which indicates that tag distribution for the resource has become stable. Similar observations have also been obtained in previous research work [9, 10]. This stable property is believed to be caused by the two public factors of social tagging, imitation and shared knowledge [9].

(2) **Consistent**, which means that collaborative annotations reflect the content of web resources, although tags may differ literally from words in the resource. On one hand, with the "wisdom of the crowds", collaborative annotations are more likely to be a less biased semantic description of web resources, which have the comparable expression capability as words in web resources. For example, we show the top 10 frequent keywords in a bookmarked resource as well as the top 10 frequent tags annotated by users in Table 1. The collaborative annotations clearly reflect what the corresponding web resource talks about. We further found that when top 10 frequent keywords considered, 73.6% of all resources are fully covered by their collaborative annotations. On the other hand, tags may differ literally from words in web resources, since they are created by large number of web users and usually represent a higher-level abstraction on the content [15]. Take the annotations shown in Table 1 as an example, beyond those matched with words, we can also observe some high-level abstraction tags, such as "programming" and "tutorial".

(3) **Shareable**, which means that most tags in collaborative annotations are shared by a large variety of web resources. We plot the average document frequency (DF) of the top N frequent tags over the data in Fig. 2. The results show that the top 10 frequent tags for web resources are shared by more than 300 resources on average. As collaborative annotations usually form a Power-Law frequency distribution, these top frequent shareable tags actually represent the major part of the collaborative annotations for web resources. This shareable property is believed to be related to two reasons: One is that users are more likely to assign similar tags for the similar content in web resources; The other is that tags which represent a semantic abstraction on the resource content are more likely to be used by a large number of users than those specific ones.

From the analysis above we can see that collaborative annotations serve as a stable and less biased semantic description for web resources. These good properties provide the foundation of our work, that is we can explore the collaboratively annotated data to perform automatic annotation. Specifically, the property (1) indicates that we can apply learning approaches on the collaboratively annotated data, while the property (2) and (3) indicate that we can predict tags for web resources by capturing the relationships between resources and annotations. Therefore, in this paper, we propose a generative topic model for automatic annotation, which learns the relationship between words and tags through their lower-dimensional representations, i.e. the hidden topics. We will explain the model in the next section.

4. GENERATIVE MODELS

We consider generative topic models that can perform the following two tasks: One is modeling the joint distribution of web resources and collaborative annotations, which can be useful for simultaneously clustering words and tags to capture low-dimensional probabilistic relationships between these two types of data; The other is modeling the conditional distribution of annotations given a particular web resource, which is useful for automatic annotation.

We focus on the collaboratively annotated data, i.e. the pair (Words, Tags). It can be formally denoted as a pair $(\boldsymbol{w}, \boldsymbol{t})$. The first element $\boldsymbol{w} = \{w_1, \ldots, w_N\}$ denotes a web resource which has a sequence of N words from a word vocabulary indexed by $\{1, \ldots, V\}$. The second element $\boldsymbol{t} = \{t_1, \ldots, t_M\}$ denotes collaborative annotations which has a collection of M tags from a tag vocabulary indexed by $\{1, \ldots, U\}$. The total collection of D annotated data then can be denoted as a corpus $\mathcal{R} = \{(\boldsymbol{w}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{w}_D, \boldsymbol{t}_D)\}$.

4.1 Word-Tag LDA

Based on the analysis above, we propose to use a generative topic model, referred as Word-Tag LDA, for automatic annotation which captures the conditional relationships between lower-dimensional representations of words and tags. Word-Tag LDA is represented as a probabilistic graphical model in Fig. 3. The model extends the conventional LDA model by simultaneously modeling the generative process of web resources and the process of collaborative annotations. In particular, it first generates words from an LDA model. Then only the topics associated with the words in the resource are used to generate the tags, resulting in a coupling between word and tag topics. Formally, let K be the latent topic number, let $\boldsymbol{z} = \{z_1, z_2, \ldots, z_N\}$ be the latent topics that generate the words, and let $\boldsymbol{y} = \{y_1, y_2, \ldots, y_M\}$ be the indexing variables that take values from 1 to N with equal probability. Word-Tag LDA assumes the following generative process of each collaboratively annotated resource $(\boldsymbol{w}, \boldsymbol{t})$ in corpus $\mathcal{R}$:

1. Draw topic proportions $\theta \sim \text{Dir}(\alpha)$
2. For each of the N words w_n
   (a) Draw topic assignment $z_n \sim \text{Multinomial}(\theta)$
   (b) Draw word $w_n \sim p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n
3. For each of the M tags t_m
   (a) Draw index $y_m \sim \text{Unif}(1, \ldots, N)$
   (b) Draw tag $t_m \sim p(t_m|y_m, \boldsymbol{z}, \eta)$, a multinomial probability conditioned on the topic z_{y_m}

Given the parameters α, β and η, we obtain the marginal distribution of a pair $(\boldsymbol{w}, \boldsymbol{t})$:

$$p(\boldsymbol{w}, \boldsymbol{t}|\alpha, \beta, \eta) = \int p(\theta|\alpha)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)) (\prod_{m=1}^{M}\sum_{y_m} p(y_m|N)p(t_m|y_m, \boldsymbol{z}, \eta))d\theta \quad (1)$$

where θ is a K-dimensional Dirichlet random variable, and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_K^{\alpha_K-1} \quad (2)$$

Finally, taking the product of the marginal probabilities of single collaboratively annotated resources, we obtain the probability of a corpus:

$$p(\mathcal{R}|\alpha, \beta, \eta) = \prod_{d=1}^{D}\int p(\theta_d|\alpha)(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)) (\prod_{m=1}^{M_d}\sum_{y_{dm}} p(y_{dm}|N_d)p(t_{dm}|y_{dm}, \boldsymbol{z}_d, \eta))d\theta_d \quad (3)$$

where notations with subscript d denote parameters/variables corresponding to the d-th annotated resource in the corpus.

The generative process of the Word-Tag LDA is essentially the same as the Correspondence LDA model proposed in [3] with the difference that the Word-Tag LDA model imitates the generation of web resources and their corresponding collaborative annotations, while [3] models the relationship between image regions and captions.

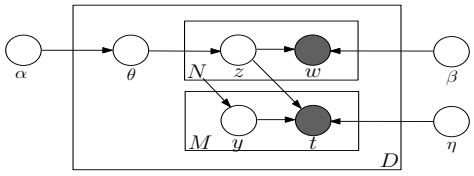

Figure 3: Graphical model of Word-Tag LDA.

Let us consider the two aforementioned tasks under our Word-Tag LDA model. For the first task, Word-Tag LDA models the joint distribution of web resources and collaborative annotations in the form of Eqn. (1). Here we take some further analysis on how Word-Tag LDA captures the probabilistic relationship between these two types of data. Word-Tag LDA models the conditional correspondence between word and tag topics, by assuming that the topic for the tag is selected uniformly out off the assignments of topics in words. This coupling between word and tag topics actually captures the fact that the resource is generated first and then annotated with tags. It also reflects how Word-Tag LDA understands the consistent property of collaboratively annotated data, that is the collaborative annotations created by many web users conceptually agree with the original authors over the topics of the web resource. Moreover, this conditional relationship between word and tag topics also ensures that the resulting topics (i.e. multinomial parameters) over words and tags will correspond.

For the second task, we can calculate the conditional probability $\Pr(t|\boldsymbol{w})$ needed for automatic annotation based on the variational inference methods. Given a web resource and Word-Tag LDA, we can compute an approximate posterior distribution over topics for each word. Using these parameters, we perform automatic annotation by finding the corresponding distribution over tags given words as follows:

$$p(t|\boldsymbol{w}) \approx \sum_{n}\sum_{z_n} q(z_n|\phi_n)p(t|z_n, \eta)$$

where $q(z_n|\phi_n)$ is the approximated posterior distribution over the topics of the n-th word.

4.2 Inference and Estimation

In this section, we describe approximate inference and parameter estimation for the Word-Tag LDA model. As a side effect of the inference method, we can compute approximations to our distribution of interest: $\Pr(t|\boldsymbol{w})$.

4.2.1 Variational Inference

Exact probabilistic inference for Word-Tag LDA is intractable as in conventional LDA. Therefore, we use variational inference method to approximate the posterior distribution over the latent variables given a particular annotated web resource.

Specifically, we define the following factorized distribution on the latent variables:

$$q(\theta, \boldsymbol{z}, \boldsymbol{y}|\gamma, \phi, \varphi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n)\prod_{m=1}^{M} q(y_m|\varphi_m)$$

with free (variational) parameters γ, ϕ, and φ, where γ is a K-dimensional Dirichlet parameter, ϕ_n are N K-dimensional multinomial parameters, and φ_m are M N-dimensional multinomial parameters.

We then minimize the KL-divergence between this factorized distribution and the true posterior by taking derivatives of the lower bound of the log likelihood function with respect to the variational parameters. Let β_{iv} be $p(w_n^v|z_n = i, \beta)$ for word v. We can obtain the following update equations:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

$$\phi_{ni} \propto \beta_{iv} \exp\Big\{\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j) + \sum_{m=1}^{M} \varphi_{mn} \log p(t_m | y_m = n, z_m = i, \eta)\Big\}$$

$$\varphi_{mn} \propto \exp\Big\{\sum_{k=1}^{K} \phi_{ni} \log p(t_m | y_m = n, z_n = i, \eta)\Big\}$$

These update equations are invoked repeatedly until the KL divergence converges.

4.2.2 Parameter Estimation

Given a corpus of collaboratively annotated data, we find maximum likelihood estimation of the model parameters with a *variational EM* procedure which maximizes the lower bound on the log likelihood of the data induced by the variational approximation described above. Specifically, the E-step computes the variational posterior for each collaboratively annotated web resource given the current setting of the parameters. The M-step subsequently finds maximum likelihood estimation of the model parameters with expected sufficient statistics under the approximate posterior distribution. These two steps are repeated until the lower bound on the log likelihood converges.

5. EXPERIMENTAL RESULTS

5.1 Data Set

For experiments, we collected a sample of Delicious data by crawling its website during October and November, 2008. The data set consists of 167,958,659 taggings made by 825,402 different users on 57,813,581 different URLs with 5,916,196 different tags. We empirically filtered out the URLs annotated by less then 100 bookmarks, as the annotations for such URLs might not have achieved stabilization for learning. This left us 904,397 URLs. We then disregarded users, merged all the tags of the same URL together, crawled web resources referred by the URLs and obtained the pairs (Words, Tags). We ended up with 71,608 pairs as the collaboratively annotated data, which include 1,513,338 distinct words and 87,302 distinct tags after normalization. In our experiments, we randomly selected 40,000 pairs as the training set and 10,000 pairs as the testing set.

5.2 Baseline Methods

To demonstrate the effectiveness of our automatic annotation approach, here we first introduce a set of representative baseline methods used in our experiments, including keyword approach, similarity approach and joint approach.

Keyword Approach

Since collaborative annotations reflect the content of web resources, keywords from the resource content are good candidates for tags. The keyword approach simply generates tags by extracting keywords from the web resources. It can be viewed as the most intuitive and basic approach for automatic annotation. Specifically, words from web resources are ranked in different ways and the top ranked words are extracted as the predicted tags. Here we ranked words based on *tf* (Key_Tf) and *tf-idf* (Key_TfIdf) respectively.

Similarity Approach

The similarity (Sim) approach annotates web resources in a collaborative filtering way. It calculates the cosine similarity between a test web resource $\boldsymbol{d}_{test}$ and each training resource $\boldsymbol{d}_i$ in the form $Sim(\boldsymbol{d}_{test}, \boldsymbol{d}_i) = \frac{\sum_j c(\boldsymbol{d}_{test}, w_j) c(\boldsymbol{d}_i, w_j)}{\sqrt{\sum_j c(\boldsymbol{d}_{test}, w_j)^2} \sqrt{\sum_j c(\boldsymbol{d}_i, w_j)^2}}$ where $c(\boldsymbol{d}, w_j)$ represents the count of the j-th word in resource $\boldsymbol{d}$. The top t tags from s most similar resources are then considered. Here we set t to be 2 and s to be 5, resulting in 10 tags for each test resource.

Joint Approach

The joint (Joint) approach also employs a topic model for automatic annotation, which is essentially the same as the mixed membership model [8]. Comparing with Word-Tag LDA, the topic model in Joint approach makes a conditional independence assumption between word and tag topics, that is the topic assignment for each word and tag is assumed to be generated conditioned on the same topic distribution respectively. Specifically, its generative process is:

1. Draw topic proportions $\theta \sim \text{Dir}(\alpha)$
2. For each of the N words w_n
   (a) Draw topic assignment $z_n \sim \text{Multinomial}(\theta)$
   (b) Draw word $w_n \sim p(w_n|z_z, \beta)$, a multinomial probability conditioned on the topic z_n
3. For each of the M tags t_m
   (a) Draw topic assignment $v_m \sim \text{Multinomial}(\theta)$
   (b) Draw tag $t_m \sim p(t_m|v_m, \eta)$, a multinomial probability conditioned on the topic v_m

In annotation, we approximate the conditional distribution over tags based on variational methods as follows:

$$p(t|\boldsymbol{w}) = \sum_v p(t|v) \int p(v|\theta) p(\theta|\gamma) d\theta$$

where γ denotes the approximate posterior Dirichlet. The integral over θ is easily computed; it is the expectation of the k-th component of $\theta \sim Dir(\gamma)$:

$$\int p(v|\theta) p(\theta|\gamma) d\theta = \frac{\gamma_v}{\sum_v \gamma_v}$$

5.3 Topics

We applied our Word-Tag LDA on the training set. Fig. 4 presents the log-likelihood on the training data by choosing different number of hidden topics and with different iteration times. In Fig. 4 we can find that the log-likelihood increases fast from 10 topics to 300 topics and slows down

Table 2: Comparisons between topics over words and tags.

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Word	Tag	Word	Tag	Word	Tag	Word	Tag	Word	Tag
secure	secure	game	game	economic	economic	hotel	travel	design	webdesign
network	network	cheat	videogame	economy	economy	flight	hotel	website	design
scan	sysadmin	wii	nintendo	market	longtail	airline	airline	web	gallery
nmap	hack	nintendo	ds	long	fax	travel	flight	blog	inspiration
tool	infosec	xbox	wii	tail	capital	airport	airfare	inspiration	css
sniff	audit	ds	cheat	fax	bubble	ticket	ticket	tutorial	showcase
detect	compute	play	playstation	price	long	air	cheap	beautiful	web
packet	exploit	pc	entertain	cost	tail	city	airplane	site	layout
system	pentest	playstation	mario	capital	interest	cheap	vacation	part	website
list	software	mario	nds	economist	dollar	airfare	fly	list	shadow

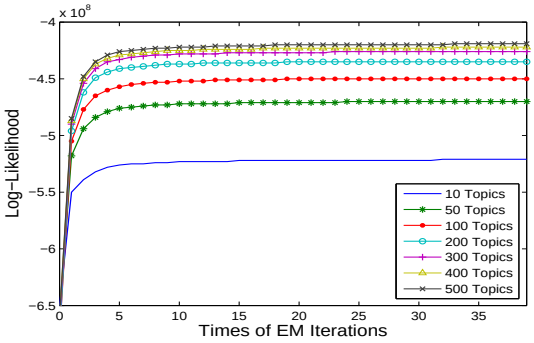

Figure 4: Log-Likelihood on the times of iteration of different number of topics with Word-Tag LDA.

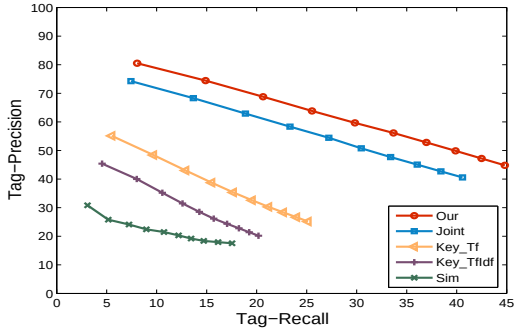

Figure 5: Precision-Recall graph shows the change of Tag-Precision and Tag-Recall with the number of predicted tags increases.

when topic number is larger than 300. It indicates that the hidden topic with 300 dimensions is basically enough to capture the major category of meanings in the training data. Higher dimensions are very probably to be redundant and may cause the problem of over-fitting. Therefore, in our experiment, we model our data with 300 topics.

Here we also show the corresponding topics over words and tags learnt in Word-Tag LDA. We choose the corresponding top 10 words and tags under each topic according to the probability $p(w|z,\beta)$ and $p(t|v,\eta)$ respectively, and randomly list 5 topics in Table 2. From Table 2, we can see that each resulting topic over words and tags corresponds to each other very well and form a special category of semantics. For example, Topic 1 is mainly about "network security", while Topic 3 talks about "economy" which is clearly more related to the "long tail" theme specifically. Moreover, we can find that words and tags under the corresponding topics may differ literally. For example, "longtail" appears in the top ranked tags under Topic 3 but not in the top ranked words. It indicates that users are likely to create such a special tag to summarize the corresponding content.

5.4 Annotation Performance

Since we are primarily interested in automatic annotation for web resources, we evaluate our approach and four baseline methods (i.e., Key_Tf, Key_TfIdf, Sim and Joint) on the annotation task. We propose to use the following metrics [21] to measure the effectiveness of our approach.

- *Top-k accuracy*: Percentage of web resources correctly annotated by *at least* one of the top kth predicted tags.
- *Exact-k accuracy*: Percentage of web resources correctly annotated by the kth predicted tag.
- *Tag-recall*: Percentage of correctly predicted tags among all tags annotated by the users.
- *Tag-precision*: Percentage of correctly predicted tags among all tags predicted by the algorithm.

Due to the space limitation, we only present the evaluation over the top 10 tags here. The topic number K in Joint approach is also set to 300 for a fair comparison.

The annotation performance in terms of top-k and exact-k accuracy on the testing set is depicted in Fig. 6. The results show that our approach can outperform all the other baselines significantly (p-value<0.01). As shown in Fig. 6(a), our approach makes 80.5% correct annotation for the top-most (top-1) tag, for which all the other baselines are below 75%. With the increase of the tag number, the top-k performance gradually improves. The accuracy of top-10 tags for our approach reaches 99.1%, indicating that at least 1 of 10 tags annotated by our approach is also annotated by the users over 9,900 test web resources. From Fig. 6(b) we can see that our approach reaches higher accuracy than the others at different ranks. The top-most tags predicted by our approach achieve the best accuracy (80.5%) and the performance decreases as the rank of the tags decreases. The precision-recall graph is depicted in Fig. 5. Again, we can see that our approach is clearly the winner.

For the two topic model based approaches, our approach and Joint approach, we made further comparison in terms of annotation perplexity [3]. Specifically, we computed the perplexity of the given annotations under $p(t|\boldsymbol{w})$ for each web resource in the testing set to measure the annotation quality. Perplexity, used by convention in language modeling, is equivalent algebraically to the inverse of the geometric mean

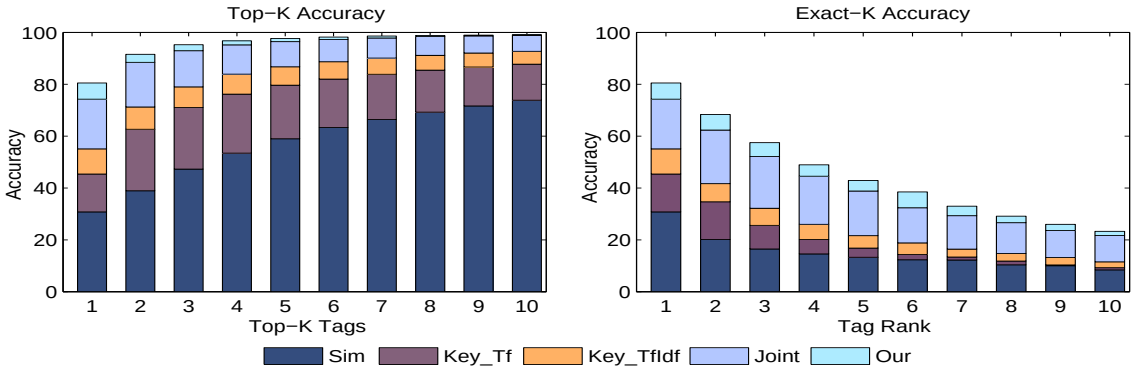

Figure 6: Annotation performance in terms of Top-K Accuracy and Exact-K Accuracy.

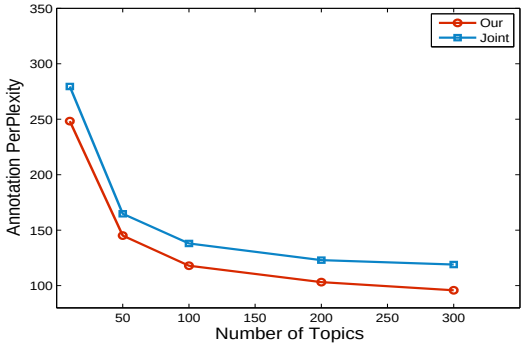

Figure 7: Annotation Perplexity on the testing set (lower numbers are better).

per-word likelihood. A lower perplexity score indicates better generalization performance. Formally, for a testing set of D annotated resources, the annotation perplexity is:

$$perplexity = exp\Big\{ - \frac{\sum_{d=1}^{D} \sum_{m=1}^{M_d} \log p(t_m|\boldsymbol{w}_d)}{\sum_{d=1}^{D} M_d} \Big\}$$

Fig. 7 shows the perplexity of the held-out annotations under the maximum likelihood estimation of each model for different values of topic number K. We see that our approach constantly outperforms Joint approach, which is consistent with the performance results above.

5.5 Case Study

Table 3 shows the top 10 user-generated tags for two sampled web resources from testing set, as well as the top 10 tags predicted by each approach. The underline indicates an overlap between the prediction and ground truth. These examples illustrate the limitations and power of all the annotation approaches.

From the examples we can see that Sim approach gives the least impressive annotation among the five. It produces more unrelated tags because such tags are used to annotate similar resources, e.g. "optimization" may annotate a similar resource about hardware but might not be appropriate for resource #2374. This is also reflected in the previous quantitative evaluation results, where Sim approach cannot achieve high performance in terms of all the measures. It seems that only leveraging the similarity between resources to predict annotations is not sufficient, since the relationship is too coarse.

The two keyword approaches perform better than Sim approach, indicating that users are likely to pick up keywords as tags for resources. As shown in previous section, the top-1 accuracy of Key_Tf approach reaches 55.1%, showing that more than half resources would be annotated by the most frequent word. Among these two keyword approaches, Key_Tf approach achieves better results than Key_TfIdf. This is because Key_TfIdf approach prefers to put more specific words at top ranks, which might not be appropriate candidates for tags, e.g. "telco" for resource #5139. Moreover, both keyword approaches have limited annotation performance as they make a strict consistent assumption between words and tags. They both suffer from ignoring the fact that tags may differ literally from words in web resources, since tags are created by large number of web users and usually represent a higher-level abstraction on the resource content, e.g. "tutorial" and "hardware" for resource #2374.

By capturing the low-dimensional probabilistic relationships between words and tags, we can improve the annotation performance, e.g. 7 of 10 predicted tags match with user-generated annotations for resource #5139 in our approach. Among these two topic model based approaches, our approach achieves better results than Joint approach. In fact, due to the decoupling of word topics and tag topics, Joint approach easily allows tags to be generated by topics that did not contribute to generating the words. In contrary, Word-Tag LDA forces a greater degree of correspondence between word and tag topics, leading to better estimation of model parameters and higher performance in automatic annotation. Besides, topic model based approaches also show some limitations as they might fail to predict very specific tags at top ranks. For example, the top 1 frequent tag "iptv" in user-generated annotations for resource #5139 does not appear in top annotations from either Joint approach or our approach, while both keyword approaches can predict it correctly. It indicates that beyond topics, we may also try to capture more direct relationship between words and tags to help eliminate such problem.

6. CONCLUSIONS

In this paper, we investigate the problem of automatic annotation, which aims to predict social tags for web resources automatically to help future navigation, filtering or search. We propose to use a probabilistic topic model, referred as Word-Tag LDA, to capture the relationship be-

Table 3: Example annotated web resources and their automatic annotations under different approaches.

Resource No.	Approach	Top 10 Tags
#2374	User	ubuntu eeepc linux howto tutorial eee install laptop asu hardware
	Sim	linux software opensource install installation css optimization javascript performance html
	Key_TF	eee ubuntu sd install card select windows installation option unetbootin
	Key_Tfldf	eee sd ubuntu unetbootin subnotebook netbook card install technet sony
	Joint	hardware linux ubuntu windows howto microsoft compute software programming reference
	Our	eeepc linux ubuntu apple mac eee tutorial hardware gadget windows
#5139	User	iptv tv television video network reference ip technology internet media
	Sim	video bookmarklet download youtube google book ebook free programming reference
	Key_TF	stream iptv video channel network box local office television system cable
	Key_Tfldf	telco iptv stream channel multicast television cable packet bandwidth vod
	Joint	network video tv programming compute technology internet oreil reference microsoft
	Our	network video tv technology compute secure internet media apple television

tween resources and annotations by statistical modeling of the collaboratively annotated data. With Word-Tag LDA learnt in hand, we are able to predict the distribution of tags for new coming web resources. Experimental results show that our approach can significantly outperform all the other baseline methods in automatic annotation.

Our future work includes investigating the usefulness of our automatic annotation approach in other applications, such as resource discovery and tag based retrieval. Besides, how to model the direct relationship between words and tags as well as their relationship through topics might also be an interesting issue in the future.

7. ACKNOWLEDGEMENT

This research work was funded by The 863 Hi-Tech Research and Development Program of China under grant number $2006AA01Z452$ and National Natural Science Foundation of China under grant number 60873245.

8. REFERENCES

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, 2007.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):28–37, 2001.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] P. A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-tag: large scale automatic generation of personalized annotation tags for the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 845–854, 2007.

[6] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. V. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *WWW '03: Proc. of the Twelfth International World Wide Web Conference*, pages 178–186, May 20-24 - Budapest, Hungary 2003.

[7] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM Trans. Web*, 1(2):7, 2007.

[8] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5220–5227. press, 2004.

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

[10] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, 2007.

[11] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[12] S. Handschuh and S. Staab. Authoring and annotation of web pages in cream. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 462–473, New York, NY, USA, 2002. ACM.

[13] S. Handschuh, S. Staab, and R. Volz. On deep annotation. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 431–438, 2003.

[14] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, 2008.

[15] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, 2008.

[16] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *ISWC 2005, LNCS 3729*, pages 522–536, 2005.

[17] N. F. Noy, M. Sintek, S. Decker, M. Crubézy, R. W. Fergerson, and M. A. Musen. Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.

[18] E. Quintarelli. Folksonomies: power to the people. ISKO Italy-UniMIB Meeting, June 2005.

[19] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 327–336, 2008.

[20] G. Smith. Folksonomy: Social classification. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html, August 2004.

[21] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 515–522, 2008.

[22] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, 2006.

[23] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 715–724, 2008.