Hierarchical Bayesian Models for Collaborative Tagging Systems

Markus Bundschus*, Shipeng Yu†, Volker Tresp‡, Achim Rettinger§, Mathaeus Dejori¶
and Hans-Peter Kriegel*
*Institute for Computer Science, Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 München, Germany
Email: {bundschu, kriegel}@dbs.ifi.lmu.de
†CAD & Knowledge Solutions, Siemens Medical Solutions, 51 Valley Stream Parkway, Malvern, PA 19355, USA
Email: shipeng.yu@siemens.com
‡Corporate Technology, Siemens AG, Otto-Hahn-Ring 6, 81739 München, Germany
Email: volker.tresp@siemens.com
§Institute for Computer Science (i7), Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany
Email: achim.rettinger@cs.tum.edu
¶Integrated Data Systems Dep., Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, USA
Email: mathaeus.dejori@siemens.com

***Abstract*—Collaborative tagging systems with user generated content have become a fundamental element of websites such as *Delicious*, *Flickr* or *CiteULike*. By sharing common knowledge, massively linked semantic data sets are generated that provide new challenges for data mining. In this paper, we reduce the data complexity in these systems by finding meaningful topics that serve to group similar users and serve to recommend tags or resources to users. We propose a well-founded probabilistic approach that can model every aspect of a collaborative tagging system. By integrating both user information and tag information into the well-known Latent Dirichlet Allocation framework, the developed models can be used to solve a number of important information extraction and retrieval tasks.**

***Keywords*-collaborative tagging; LDA; user modeling;**

I. Introduction

Collaborative knowledge platforms have recently emerged as popular frameworks for sharing information between users with common interests. Some popular examples of such systems are Delicious, CiteULike[1] or Flickr. A key feature of these systems is that large numbers of users upload certain resources of interest and label them with personalized tags. The resources are in most cases some type of high-dimensional data such as text documents or images. Without further processing, those resources do not contain any semantic information that is usable for automated analysis. However, meaningful annotations adding semantic to the raw resources are also given in the form of user specified tags. In contrast to taxonomies, where labels represent ordered predefined categories, no restrictions apply to tags, which are flat and chosen arbitrarily. These free-form strings actually serve the purpose to organize the resources of one single specific user. Tags might be polysemous and different users use slightly different variations of tags to express the same semantics (e. g. consider the tags *information_retrieval*, *information-retrieval* and *IR*). Also the meaning of a particular tag, such as *to_read*, might be subjective to individuals and does not necessarily express the same shared semantic for the whole community. These aspects make the extraction of meaningful information from collaborative systems both challenging and rewarding.

In this paper, we present a unified probabilistic framework for collaborative tagging systems, which has a sound theoretical foundation in Hierarchical Bayesian Statistics. By extending one well established model for document collections, the *Latent Dirichlet Allocation* (LDA) model, we are able to exploit the complete spectrum of information available in collaborative tagging systems. Hereby, all involved entities, i. e. the users, their resources and the assigned resource tags are modeled by a latent multinomial topic distribution. With this strategy, we map each entity into a common lower dimensional latent topic space and thus are able to extract structure and drastically reduce the great variety of ambiguous information inherent in collaborative tagging systems. The here proposed models can be applied naturally to various tasks. We present results for the extraction of statistical relationships between users, resources and tags. As a quantitative evaluation, we present results on assessing user similarities, a perplexity analysis on tag annotation quality, and results on personalized tag recommendation. In the latter case, we outperform several standard tag recommendation algorithms. We train our models on a fraction of the CiteULike system. CiteULike is a system that allows researchers to manage their scientific reference articles. It tries to help scientists to cope with the increasing interdependent topical complexity of today's research. While in this work, we focus on collaborative tagging systems based on text, the described models are general and could handle various types of resources such as pictures as well.

The outline of the paper is as follows: In Section II we briefly summarize existing related work. Section III

[1] http://www.citeulike.org/

introduces the hierarchical Bayesian models for collaborative tagging systems and Section IV describes our experimental setup in detail and presents results. Concluding statements as well as an outlook for future work is given in Section V.

II. Related Work

Collaborative Tagging Systems Research on collaborative tagging systems, also referred to as folksonomies or social bookmarking systems, is a relatively new research area. Here, the most popular application is tag recommendation. Some type of collaborative filtering techniques are often applied to this problem [1] or some type of machine learning algorithm such as Support Vector Machines are used for prediction of the most popular tags [2]. Typically, these algorithms are applied on a "dense" fraction of resources and tags, i. e. the resources and tags have to co-occur a sufficient number of times. [1] presents an algorithm *FolkRank*, which is based on the original PageRank algorithm for ranking web sites. So far, the recommendation algorithms exploit either the provided information from the entire community or the graph structure of the folksonomy to make predictions. In content-based recommendation algorithms, tags are derived from an analysis of the content of the resource. But tag recommendation is only one of many interesting tasks in these complex systems. Information retrieval issues [3] , the extraction of statistical relations between involved entities in the folksonomy and its mapping to taxonomies [4] as well as knowledge acquisition [5] are also of particular interest.

Our contribution provides an integrated view on the just outlined work and applications. Since we define a unified probabilistic model for collaborative tagging systems, we can apply our models in very different scenarios and tasks (see Section IV).

Probabilistic Topic Models In this area, powerful techniques such as Latent Dirichlet Allocation (LDA) [6] have been proposed for the automated extraction of useful information from large document collections. In its classical application, LDA tries to find the underlying latent semantic properties in an unsupervised fashion. Depending on the addressed generative process, various extensions of the LDA framework have been proposed (see e. g. [7], [8], [9]).

III. Models

A. Terminology

Entities in a social tagging system consist of finite sets of users U, resources R and $Tags$. Following the notion of [3], a social tagging system or folksonomy F can be represented as a four-tuple:

$$F = \langle U, R, Tags, P \rangle, \tag{1}$$

where $P \subseteq U \times R \times Tags$ denotes a ternary relation. Each post p can be represented as a triple:

$$p \subseteq \{\langle u, r, T_{ur} \rangle : u \in U, r \in R, T_{ur} \in T_u\}. \tag{2}$$

Note that $T_{ur} \subseteq T_u \subseteq Tags$ and T_u represent the set of tags for a specific user u. A tag label l denotes a specific tag from T_u.

B. Classical LDA

In LDA, the generation of a resource collection is modeled as a three step process. First, for each resource r, a distribution over topics is sampled from a Dirichlet distribution. Second, for each word w in the resource r, a single topic is chosen accordingly. Finally, a word is sampled from a multinomial distribution over words specific to the sampled topic. The hierarchical Bayesian model shown in Figure 1 (left) depicts this generative process. α and β denote symmetric Dirichlet priors. θ represents the resource specific multinomial distribution over T topics, each being drawn independently from α. Φ denotes the multinomial distribution over N vocabulary items for each of T topics being drawn independently from β. For each of the N_r words w in resource r, z denotes the topic assignment for w, drawn from θ. w is drawn from the topic distribution Φ conditioned on z.

C. Topic-Tag (TT) Model

The first proposed model aims at exploiting the additional information inherent in the user specified tags of collaborative tagging systems. It extends the LDA framework by simultaneously modeling the process of generating a resource r and the process of indexing a resource with tags. In addition to the steps mentioned in the section above, two further steps are introduced (see Figure 1 (middle)). For each of the M_r tags in the resource, a topic $\tilde{z}$ is uniformly drawn based on the topic assignments for each word in the resource. Finally, each tag label l is sampled from a multinomial distribution over tag labels specific to the sampled topic. The topic assignment $\tilde{z}$ is selected uniformly out off the assignments of topics from the N_r words in resource r. Thus, we first sample an index i from $\mathrm{Uniform}(1, 2, \ldots, N_r)$ and then use the topic assignment from the word with index i to sample a tag label l_j from tag-topic distribution $\Gamma_{\tilde{z}_i}$. This leads to a coupling between both generative components. Thus, the tags of the resource are conditioned on the factors, which are present in r. This model captures the notion of first generating the content of the resource and than the tags which annotate the resource. The principle idea of coupling Θ and Γ has previously been applied successfully to modeling images and their captions [6]. There, this model outperformed several other generative annotation models. In this model, Γ denotes the vector of multinomial distribution over M_r tags for each of T topics being drawn independently from a symmetric Dirichlet prior γ. After the generation of words, a topic $\tilde{z}$ is drawn from the resource specific distribution, and a label l is drawn from the $\tilde{z}$ specific distribution Γ. Instead of estimating the parameters directly

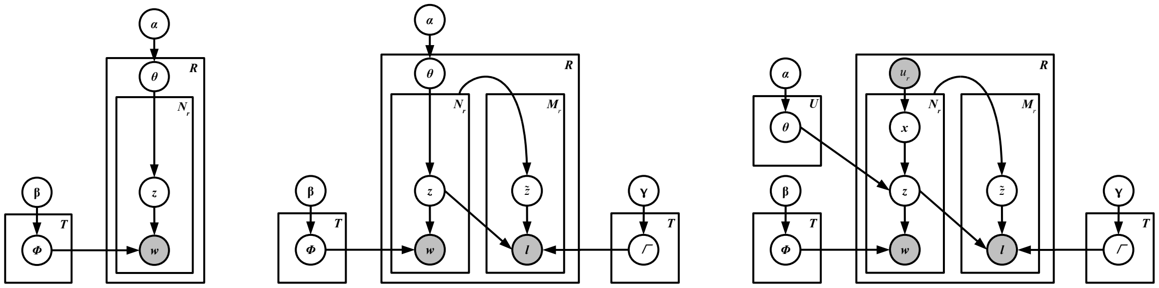

Figure 1. Graphical models in plate notation with observed (gray circles) and latent variables (white circles). Left: standard LDA. Middle: Topic-Tag (TT) model. Right: User-Topic-Tag (UTT) model

[8], we estimate Φ, θ and Γ from the posterior distributions via Gibbs Sampling [7].

D. User-Topic-Tag (UTT) Model

In this section we introduce the most elaborate model for collaborative tagging systems. The UTT model adds an additional layer accounting for the most essential entity: the user, who assigns one or more tags to web resources. This can be formalized by a two-step process where a user first cites an article based on his interests and afterwards assigns tags based on the content of the resource. This process can be modeled by a hierarchical generative model, in which each word w of a resource r is associated with two variables: a user u and a latent topic variable t. We assume that each user is interested in several topics, thus each user has a multinomial distribution over topics. First, a user u is chosen uniformly at random for each word of a certain resource. Hereby u is chosen from the users U_r, the users which cite the resource r. Second, a topic is sampled for each word from the user specific topic distribution Θ_u from user u_x chosen for that word. Third, for each of the tags associated with the resource, a topic is uniformly drawn based on the topic assignments for each word in the web page (Figure 1, right). This can be summarized as:

1) For each user $u = 1 \dots |U|$ choose $\Theta_u \sim Dirichlet(\alpha)$
   For each topic $t = 1 \dots |T|$ choose $\Phi_t \sim Dirichlet(\beta)$ and $\Gamma_t \sim Dirichlet(\gamma)$

2) For each resource $r = 1 \dots |R|$ and its given users U_r
   For each word w_i, $i = 1 \dots N_r$ in resource r
   a) Sample a user $x_i \sim Uniform(1, \dots, U_r)$
   b) Given x_i, sample a topic $z_i \sim Mult(\Theta_{u_i})$
   c) Given z_i, sample a word $w_i \sim Mult(\Phi_{z_i})$

3) For each tag label l_j, $j = 1 \dots M_r$ in resource r
   a) Sample an index $i \sim Uniform(1, \dots, N_r)$
   b) Given topic $\tilde{z_i}$, sample a tag label $l_j \sim Mult(\Gamma_{\tilde{z_i}})$

For the sake of brevity, we omit the Gibbs Sampling equations and provide them in an extended version of this paper online[2]. In the UTT model, the interest of the user is modeled by the assignments of users to words in the resource. Obviously, this is a simplifying modeling assumption. However, this assumption yielded promising results in the past when modeling authors and their interests [7]. Furthermore, once we have trained a UTT model, we can estimate the resource specific topic distribution based on a single user. This provides a personalized view on a resource and results in a potential better tag recommendation (see Section IV-B4).

Table I
CORPUS STATISTICS CITEULIKE DATA SET

	Unique	Total
Resources	18.628	18.628
Word tokens	14.489	1.161.794
Tags	4.311	125.808
Users	1.393	18.628

IV. EXPERIMENTS

A. Experimental Setup

Data Set: CiteULike provides data snapshots on their webpage[3]. The data used in our experiments was from November 13th 2008.

Training Data: We selected a reasonable high number of users (1393) and included articles that were cited by at least three users. Word tokens from title and abstract were stemmed with a standard Porter stemmer and stop words were removed. Word tokens and tags occurring less than five times were filtered out. Table I summarizes the corpus statistics. The user id's, resource id's and tags are provided as supplementary data[4]. All in all, this results in a total number of 64159 posts.

Test Set for Tag Recommendation: The only restriction for the test set was that a resource had to be posted from a user previously seen in the training set. The same applies to tags. We evaluate the models on a total of 15000 posts. In average each user uses 32 tags. The maximum number of tag labels for a specific user is 279. The average number of tag assignments for a user is three.

[2] www.dbs.ifi.lmu.de/~bundschu

[3] http://www.citeulike.org/faq/data.adp

[4] www.dbs.ifi.lmu.de/~bundschu/UTTmodel_supplementary/info.html

Training Details: Parameters were estimated by averaging samples from ten randomly-seeded runs, each running over 100 iterations, with an initial burn-in phase of 500 for the TT model and 1500 iterations for the UTT model. We found the number of burn-in iterations to be a convenient choice by observing a flattening of the log likelihood. Instead of estimating the hyperparameters α, β and γ, we fix them to 50/T, 0.001 and 1/M respectively in each of the experiments (M represents the total number of tags). The values were chosen according to [7]. We trained the topic models with a predefined number of topics ranging from T=200, T=300 and T=400 to show that the performance is not very sensitive to this parameter as long as the number of topics is reasonably high. In addition, models with T=10 , T=50 and T=100 were trained for the perplexity evaluation in Section IV-B2.

B. Results

1) Uncovering the Hidden Semantic Properties: Table II illustrates two different topics (out of 200) from the corpus. Note that the coupling between $p(w|t)$ and $p(tag|t)$ is a property of the here proposed models and originates from the sampling of a topic for a specific tag based on the topic assignments of the resource (see Section III). To show the descriptive power of our learned model, we chose two topics describing different aspects of CiteULike. Topic 18 is about the science of networks, while topic 84 reflects a topic about information retrieval. All extracted 200 topics from the TT and UTT model are available as supplementary data [4].

Information about $p(u|t)$ in CiteULike provides interesting insights about the main research interests of users. The most likely users given the topics for a UTT model with T=200 are again provided as supplementary data[4].

Other interesting statistical relationships that can be extracted with the here derived models are e. g. the statistical relationships between tag labels and topics $p(t|l)$. $p(t|l)$ gives us a notion about the involvement of a tag in different topics. Table III shows one such example for the tag *semantic*. *semantic* is mostly discussed in the context of the traditional semantic web, but also in the context of bioinformatics and web services. The third topic discusses the tag in the classical information retrieval domain.

2) Tag Perplexity: In addition to the qualitative evaluation of the TT and UTT model shown above, we measure the tag annotation quality in terms of perplexity. Intuitively speaking, perplexity is the ability to predict tags on new unseen documents. Perplexity, a quantitative measure for comparing language models is widely used to compare the predictive performance of topic models (see e. g. [7]) and is defined over a test set as:

$$Perp.(\mathbf{l_{test}}|D_{train}) = exp\left(-\frac{\sum_{i=1}^{D_{test}} log(p(\mathbf{l_d}|D_{train}))}{\sum_{i=1}^{D_{test}} L_d}\right),$$

where $\mathbf{l_{test}}$ are the tags in the test set and $\mathbf{l_d}$ represent the tags in a certain test resource. D_{train} represents the trained parameters which differ dependent of the model (see Section III). We partition the data set into disjoint training (90%)and test sets (10%) and select for each resource in the test set a subset of 50% of the tags for the evaluation. The remaining 50% of the tags are used by standard LDA to estimate Θ, since LDA is not able to model the dependency between the tokens in the resource and the tags. In contrast, the TT and UTT model first estimate the resource specific Θ, which is estimated online via Gibbs Sampling respectively. Afterwards the most likely tags are computed by Γ. All perplexity values were computed by averaging over ten different samples. Figure 2 plots the perplexity over the held-out tags under the maximum likelihood estimates of each model for different values of T. Note that a lower perplexity indicates a better annotation quality. We see that the two models, which include the resource tokens into the computation of the likelihood clearly outperform the standard LDA model. As T increases, the UTT model has a better perplexity than the TT model (with a crosspoint at T=100). With T=400 the perplexity of the TT model starts slightly to increase, while for the UTT model the perplexity remains constant.

Table III
THREE MOST PROBABLE TOPICS FOR THE TAG **SEMANTIC**. UTT MODEL, T=200.

Topic	Topic description (word stems)
Topic 96 (p=0.85)	semant ontolog annot knowledg web commun support integr describ xml
Topic 184 (p=0.05)	servic web workflow bioinformat applic resourc integr manag standard interfac
Topic 84 (p=0.03)	search retriev inform relev ar rank feedback effect document

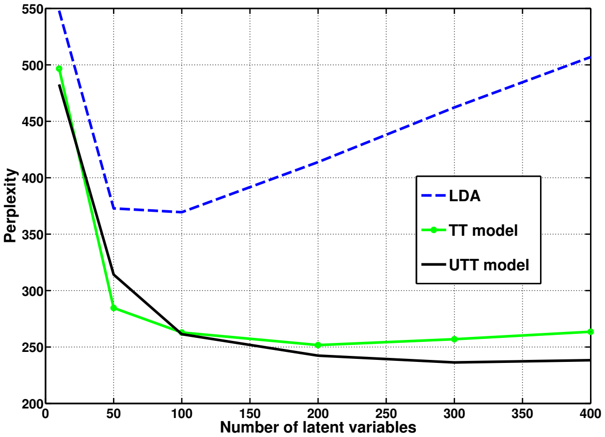

Figure 2. Tag perplexity on the test set.

Table II
SELECTED TOPICS, LEARNED FROM THE UTT MODEL WITH T=200. FOR EACH TOPIC THE FIVE MOST PROBABLY WORDS AND TAGS ARE LISTED

Topic 18			
Word	Prob.	Tag	Prob.
network	0.51	network	0.36
connect	0.040	networks	0.266
complex	0.035	graph	0.009
structur	0.023	complexity	0.009
topolog	0.020	complex	0.007

Topic 84			
Word	Prob.	Tag	Prob.
search	0.171	ir	0.21
retriev	0.125	search	0.059
inform	0.077	information-retrieval	0.054
relev	0.069	retrieval	0.042
feedback	0.029	evaluation	0.041

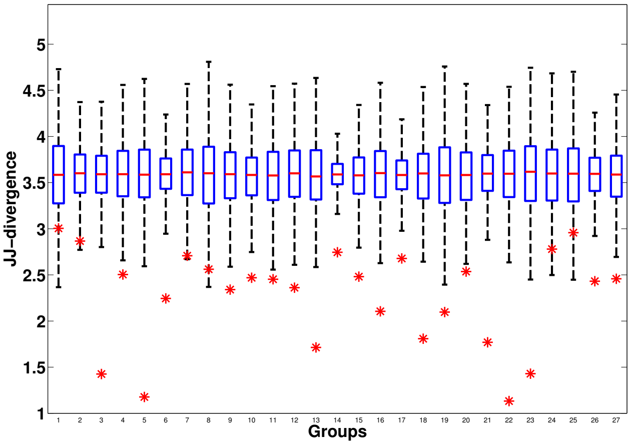

Figure 3. Boxplot over 1000 random samplings for each group. The stars indicate the true group divergence.

3) Assessing User Similarity with the UTT Model: In order to test if the UTT model is able to identify similar users, we identified all users given in our data set which are members of groups in the CiteULike system. CiteULike-groups typically share similar research interests and often belong to one research lab, like for instance the Carnegie Mellon University Human Interaction Institute with a group of 26 users. In our data set there are 488 users out of 1393 which belong to a total of 524 groups (as of November 18, 2008). We excluded all groups with less than five members. This resulted in a total of 27 groups with 160 users. 31 user belong to more than one group and the maximum number of groups for one user is five. We derive the similarity between users based on the learned user-topic distributions Θ_u. Since each user is represented as a multinomial over the topics T, Jeffreys' J-divergence a symmetric version of the Kullback-Leibler (KL) divergence, is used. Jeffreys'J-divergence originates from information theory and is a method to compute the similarity between two probability distributions.

Our assumption is that users that share the same group membership should be significantly more similar to each other than users that are randomly chosen and considered as an artificial group. Therefore, we repeated the following procedure for each group: We randomly sampled n users (with n, the size of a group) and computed the mean divergence of this artificial group. This step was repeated 1000 times. Afterwards these results are compared to the true group divergence. Figure 3 shows the corresponding boxplot for the 1000 samplings for each group. On each box, the central red line is the median, the edges of the box are the 25th and 75th percentiles. The whiskers were chosen such that all data points within $\pm 2.7\sigma$ are considered not as outliers. The stars in the plot indicate the true divergence for each group. All true group divergences fall clearly below the just mentioned percentiles. Furthermore, 20 out of 27 groups are not within $\pm 2.7\sigma$. When using the document or tag distribution of a user as baseline to compute user similarities, none of the 27 true group divergences fall out of $\pm 2.7\sigma$ (Results not shown for the sake of brevity, but available online[4]).

Table IV
NDCG EVALUATION FOR DIFFERENT NUMBER OF RECOMMENDATIONS.

NDCG	@5	@10	@15	all
Baseline 1	0.04	0.05	0.06	0.19
Baseline 2	0.14	0.20	0.23	0.37
Baseline 3	0.25	0.31	0.34	0.40
TT, T=200	0.29	0.37	0.39	0.47
TT, T=300	0.29	0.37	0.39	0.47
TT, T=400	0.30	0.37	0.40	0.49
UTT, T=200	0.31	0.37	0.40	0.50
UTT, T=300	0.32	0.38	0.41	0.51
UTT, T=400	0.34	0.40	0.43	0.52

4) Personalized Tag Recommendation: We perform evaluation on a post basis, i. e. given an user $u \in U$ and a resource $r \in R$, we want to predict a recommendation or ranking of tags $t \in T_u$.

Baselines: We follow the baseline methods of previous work on tag prediction [1], but in addition provide personalized versions. The TT and the UTT model are benchmarked against three standard tag recommendation methods.

- *Most popular tags:* Tags for a resource r are predicted based on the relative frequency in the collaborative tag system. (Baseline 1)
- *Most popular tags with user restriction:* Tags for a resource r are first ranked according to the relative frequency and than are reduced to the set of tags T_u.

(Baseline 2)

- *Most popular tags with respect to the user:* All tags $t \in T_u$ are ranked according to the relative frequency. (Baseline 3)

Tag prediction with the TT and UTT model: In the TT model, the prediction of tags for unseen documents can be formulated as follows: Based on the word-topic and tag-topic count matrices learned from the independent training data set, the likelihood of a tag label $l \in T_u$ given the test resource r is $p(l|r) = \sum_t p(l|t)p(t|r)$. The first probability in the sum, $p(l|t)$, is given by the learned topic-tag distribution. The mixture of topics $p(t|r)$ for the resource has to be estimated online. For each resource r, we independently sample topics for a small number of iterations (we used i=5) by using the word counts in Φ from the training corpus.

The likelihood of a tag label $l \in T_u$ in the UTT model is given by $p(l|r,u) = \sum_t p(l|t)p(t|r,u)$. Again, $p(t|r,u)$ has to be estimated online. Here the mixture of topics for the resource is restricted with respect to the user, i. e. we estimate the topic-distribution for the resource based on the user specific topic-distribution. Recall that every post originates from a single user, therefore the estimated topic distribution for the resource under consideration is based on this user. This estimation gives a personalized view on r and thus influences the topic distribution of the resource.

Evaluation measure: We are interested in the ranking quality of predicted tags. Here we use the normalized discounted cumulative gain (NDCG) [10] to evaluate a predicted ranking, which is calculated by summing over all the "gains" along the rank list with a log discount factor as NDCG$(\hat{R}) = Z \sum_k (2^{r(k)} - 1)/\log(1+k)$, where $r(k)$ denotes the target label for the k-th ranked item in $\hat{R}$, and Z is chosen such that a perfect ranking obtains value 1. To focus more on the top-ranked items, we also consider the NDCG@n which only counts the top n items in the rank list. In addition to the ranking scenario, we report F-measure values averaged over the users as proposed in previous work [1] in an extended version of this paper[2].

Tag Prediction Results: Table IV presents the NDCG scores. The first baseline method performs quite poor, since this model does not take into account which tags a certain user has posted so far. All other methods, i. e. Baseline 2, Baseline 3, the TT and the UTT model take this information into account. The two hierarchical Bayesian models clearly outperform all three baseline methods. Therefore, taking into account the textual resources clearly adds a benefit. The hierarchical Bayesian models are both not very sensitive to the predefined number of topics T, but a slight performance increase can be observed with an increasing number of topics.

A major advantage of the UTT model can be observed when a resource has only a title and no abstract (1223 out of 15000 posts). Since the number of observed words drastically reduces, it becomes more difficult to estimate the resource specific Θ reliable. Here, the NDCG for the TT model decreases significantly (NDCG all is 0.42 for T=200). The UTT model, in contrast, can make use of the user specific topic distribution to estimate $p(t|r,u)$ more reliably and the NDCG only decreases slightly (NDCG all is 0.48 for T=200).

V. Conclusion and Outlook

In this paper, we presented hierarchical Bayesian models for mining and modelling large systems with user generated content and massive annotation. To demonstrate its performance, we trained the model on a large fraction of the CiteULike data base. As a quantitative result we showed that the here proposed models provide a better tag annotation quality in terms of perplexity compared to the standard LDA framework. With the UTT model, we are able to create a personalized view on a resource by sampling the resource specific topic distribution through the user specific topic distribution, which we see as the reason for the performance increase in the tag recommendation task. Parts of future work will aim at investigating more ways to model users within the LDA framework.

References

[1] R. Jäschke *et al.*, "Tag recommendations in folksonomies," in *Knowledge Discovery in Databases: PKDD 2007*, 2007.

[2] P. Heymann *et al.*, "Social tag prediction," in *Proceedings of the 31st Annual International ACM SIGIR Conference*, 2008.

[3] A. Hotho *et al.*, "Information retrieval in folksonomies: Search and ranking," in *The Semantic Web: Research and Applications*, 2006.

[4] C. Cattuto *et al.*, "Semantic grounding of tag relatedness in social bookmarking systems," in *Proceedings of the 7th International Semantic Web Conference*, 2008.

[5] C. Schmitz *et al.*, "Mining association rules in folksonomies," in *Data Science and Classification*, 2006.

[6] D. M. Blei *et al.*, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, January 2003.

[7] M. Steyvers *et al.*, "Probabilistic author-topic models for information discovery," in *Proceedings of the 10th ACM SIGKDD International Conference*, 2004.

[8] D. M. Blei *et al.*, "Modeling annotated data," in *Proceedings of the 26th Annual International ACM SIGIR Conference*, 2003.

[9] R. Nallapati *et al.*, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD International Conference*, 2008.

[10] K. Jarvelin and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd annual international ACM SIGIR Conference*, 2000.