Semantic Grounding of Hybridization for Tag Recommendation

Yan'an Jin[1,2], Ruixuan Li[1,⋆], Yi Cai[3], Qing Li[3], Ali Daud[4], and Yuhua Li[1]

[1] College of Computer Science and Technology, Huazhong University of
Science and Technology, Wuhan 430074, China
jin.yan.an@smail.hust.edu.cn, rxli@hust.edu.cn, idcliyuhua@hust.edu.cn
[2] School of Information and Management, Hubei University of Economics,
Wuhan 430205, China
[3] Department of Computer Science, City University of Hong Kong, Hong Kong, China
yicai3@cityu.edu.hk, itqli@cityu.edu.hk
[4] Department of Computer Science and Technology, Tsinghua University,
Beijing 100084 ,China
ali_msdb@hotmail.com

Abstract. Tag recommendation for new resources is one of the most important issues discussed recently. Many existing approaches ignore text semantics and can not recommend new tags which are not in the training dataset (e.g., FolkRank). Some exceptional semantic approaches use a probabilistic latent semantic method to recommend tags in terms of topic knowledge (e.g., ACT model). However, they do not perform well because many entities in these models result in much noise. In this paper, we propose hybrid approaches in folksonomy to challenge these problems. Hybrid approaches are combination of Language Model (LM) for keyword based approach and Latent Dirichlet Allocation (LDA), Tag-Topic (TT) model and User-Tag-Topic (UTT) model for topic based approaches. Our approaches can recommend meaningful tags and can be used to discover resource implicit correlations. Experimental results on Bibsonomy dataset show that LM performs better than all other hybrid and non-hybrid approaches. Also the hybrid approaches with less number of entities (e.g., LDA with only one entity) perform better than those approaches having more entities (e.g., UTT with three entities) for tag recommendation task.

1 Introduction

As a promptly emerged popular way, social tags allow users to annotate, categorize and find web resources by assigning one or more descriptive words (called tags) in folksonomies, such as Flickr[1], Bibsonomy[2]. Thus, the tags assigned to each resource create a bridge between heterogeneous data and users who are accustomed to keyword-based search and browsing. To create this association, social tagging systems demand users to manually annotate tags for each resource [1]. However, manual assignment of

[⋆] Corresponding author.
[1] http://www.flickr.com
[2] http://www.bibsonomy.org

L. Chen et al. (Eds.): WAIM 2010, LNCS 6184, pp. 139–150, 2010.

tags results in the following problems: 1) Tags assigned by users are sometime very specific and sometime very general; 2) There may exist cold start problem for new resources, as new resources do not acquire enough tags quickly; 3) Tag assignment is potentially time-consuming. Therefore, many researchers try to work on proposing some tag recommendation methods for addressing the above problems [2,3,4].

In current folksonomy services, there are four major frameworks used for tag recommendation: a) Graph connectivity based approaches, e.g., FolkRank [5,3], a variation of PageRank; b) Keywords matching based collaborative filtering approaches like [4,6,7] by using similar tags and contents of similar resources; c) topic layer (semantically related cluster of words) based approaches [8]; d) Mixture of many existing approaches and additional context component [9]. The first approach only works on a dense core of folksonomy while folksonomy usually has a sparse structure. The second approach is an essential method of keywords extraction based on word frequency but not on word semantic. Both of them can only recommend tags with highest scores from the collection of tags which have been posted previously, and can not recommend tags for new resources. The third approach is suitable for sparse kind of data and can recommend tags for newly arriving resources, but it still does not perform well because it simultaneously models many entities such as words, tags and users. What is more, it creates much noise [10]. The forth hybrid approach can do better recommendations to some extent, but its applications are limited to many context factors.

Previously, we investigate tag recommendation on Bibsonomy dataset[3] by hybrid approaches of keyword and semantic (i.e., topic-related) knowledge. Keyword based approaches are precise presentation of words, while topic modeling approaches may be not. Intuitively, topic modeling approaches represent a combination of related words as a topic. Only using topic modeling itself is too loutish for tag recommendation [11]. In this paper, we individually use keyword based and semantic based approaches as baselines to predict tags at first. Then we propose hybrid approaches to combine them pairwise. For keyword based approach, we use Language Model (LM). For semantic based approach, we employ Latent Dirichlet Allocation (LDA) [12], Tag-Topic model (TT, which is adapted from Author Topic model [13]) and User-Tag-Topic model (UTT, which is adapted from Author-Conference-Topic model [14]), respectively. The experimental result shows that LM can perform better than other approaches, and hybrid approaches with fewer entities outperform the hybrid approaches with more entities.

The novelty of this work lies in: 1) We propose hybrid approaches which combine keyword based approach and topic layer based approach to deal with the tag recommendation issue; 2) We use both description words and tags associated with a resource as the candidate set individually in hybrid approach; 3) We verify the effectiveness of our approach through experiments conducted on a real-world dataset. To the best of our knowledge, we are the first to compare the effectiveness of tag recommendation approaches by using both words and tags in hybrid approaches.

The rest of the paper is organized as follows. Section 2 provides related work. Section 3 illustrates baselines and the proposed hybrid approaches for tag recommendation. In Section 4, experiments and performance evaluation about the proposed approaches are conducted. Section 5 brings this paper to the conclusions.

[3] The dataset include two parts: tags and descriptive words associated with a resource.

2 Related Work

With the emergence of social tagging systems, tag recommendation has become an attractive area of research. Some important characteristics of social tagging systems are studied. 1) Keywords appeared in a resource are usually selected as tags by users; 2) Similar kind of tags is assigned to similar kind of resources; 3) Users with the similar interests influence each other's annotation activities [15,16].

Usually, tags are recommended on the basis of tags posted to similar posts. AutoTag [4] and TagAssist [6] use some information retrieval skills to recommend tags for web blog posts. Co-occurrence of tags was used to complement user-defined tags of photographs in Flickr [17]. An association rule mining approach is proposed for tag recommendation utilizing the association structure of tags [18]. Recently, FolkRank, an adaption of famous PageRank [10], is proposed to recommend tags on the basis of graph connectivity. Although aforementioned methods are always seen in most current social tagging systems, they are unable to perform well because of not sharing tags in the candidate set.

In addition, tag recommendation approaches based on posted tags are always doubted by the requirements of sufficient annotations for a target resource and have cold-start problem for newly arriving resources, which are highlighted by Hsu and Chen [2]. Therefore, they propose a supervised tag recommendation model to predict the stabilized tag for a target resource. Meanwhile, as a major task of RSDC2008 challenge[4], tag recommendation using semantic information is proposed. In that challenge, Zhang et al [8] propose a hybrid approach based on LM and ACT model using content information to label a new resource. Tags and users of resources are modeled together by using the topic layer to capture the semantics of text. Their main assumption is that words are responsible for generating tags and users of those tags do not match with the real world situation in which users usually generate tags for resources. As shown in [10], modeling more objects together in topic layer based approaches results in poor performance. Intuitively, we propose simpler hybrid approaches by using just description words or tags that can perform better than complex topic modeling based non-hybrid approaches.

3 Hybridization for Tag Recommendation

3.1 Keyword Based Approaches

LM is widely used in natural language processing applications such as speech recognition and information retrieval. Its goal is to try to characterize, capture, and exploit regularities in the collection of documents, and assign a reasonable probability distribution to all words. Accordingly, it can score all the words in the candidate set and recommends words with highest scores for tags.

The basic idea of LM is to interpret the relevance between a document and a query as a generative probability for the query from the document model:

$$P_{LM}(q|d) = \prod_{w \in q} P(w|d) \tag{1}$$

[4] http://www.kde.cs.uni-kassel.de/ws/rsdc08: a ECML PKDD Discovery Challenge.

where *w* is a query word token in *q*. *P(q|d)* is the probability of the document model generating the query words under the Bag of Words (BoW) assumption[19]. *P(w|d)* is the probability for generating word *w* from the document model *d*, which is usually specified by using the smoothing techniques (like Dirichlet smoothing [20]). The probability is a score, calculated by following Eq.2:

$$P(w|d) = \frac{N_d}{N_d + \lambda}\frac{tf(\dot{w}, d)}{|d|} + (1 - \frac{N_d}{N_d + \lambda})\frac{tf(\dot{w}, \mathbf{D})}{|\mathbf{D}|} \tag{2}$$

where |*d*| is the length of document *d*, *tf(w,d)* is the word frequency (i.e. number of words) of word *w* in *d*, |**D**| is the number of word tokens in the whole collection, and *tf(w,**D**)* is the word frequency of word *w* in the whole collection **D**. λ is the Dirichlet prior and its value is set according to the average document length in the document collection, which is set to 10 in our case.

In this paper, we propose three keyword based tag recommendation approaches in terms of different candidate sets: Word Language Model (**wLM**), Tag Language Model (**tLM**) and **CombLM**, which are variants of Language Model.

Word Language Model (wLM). **wLM** uses description words posted by all users for a specific resource as the candidate set. In **wLM**, a resource is viewed as a composition of the description words written by all tagging users. Symbolically, for a resource *r* we can write it as: $r = \{u_1w_1 + u_2w_2 + \ldots + u_iw_i\}$, where u_iw_i is the description written by a user u_i on a specific resource. According to original LM, the probability for generating a word w_i from the resource *r* is:

$$P_{wLM}(w|r) = \frac{N_r}{N_r + \lambda}\frac{tf(\dot{w}, r)}{|r|} + (1 - \frac{N_r}{N_r + \lambda})\frac{tf(\dot{w}, \mathbf{R})}{|\mathbf{R}|} \tag{3}$$

Tag Language Model (tLM). **tLM** uses tags posted by all users for a specific resource as the candidate set. In **tLM**, a resource is viewed as a collection of the tags for a specific resource given by all users. Symbolically, for a resource *r* we can write it as: $r = \{u_1t_1 + u_2t_2 + \ldots + u_it_i\}$, where u_it_i is the tag posted by a user u_i about a specific resource *r*. Like as **wLM**, we calculate the probability for a tag t_i from the collection by Eq.4:

$$P_{tLM}(t|r) = \frac{N_r}{N_r + \lambda}\frac{tf(\dot{t}, r)}{|r|} + (1 - \frac{N_r}{N_r + \lambda})\frac{tf(\dot{t}, \mathbf{R})}{|\mathbf{R}|} \tag{4}$$

CombLM. In fact, there are both description words and tags for a given resource in a social tagging system. Description words can be considered as summary for a resource, and tags can be thought as semantic annotation for a resource. Consequently, both can be used as the candidate set for tag recommendation. We propose a simple approach for tag recommendation combining description words and tags on **wLM** and **tLM**. The score of a tag for a given resource by **CombLM** can be calculated by Eq.5:

$$P_{CombLM}(t|r) = P_{wLM}(w|r)P_{tLM}(t|r) \tag{5}$$

3.2 Topic Modeling Approaches

Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. Now, there are many variants or extensions of topic model. In this paper, we use four adaptive topic models for our task: wLDA (LDA over description), tLDA (LDA over tags), TT model [13], UTT model [14].

As a basic topic model, LDA is widely used to model documents. It specifies a simple probabilistic procedure by which documents can be generated. First, for each document d, a multinomial distribution θ_d over topics is randomly sampled from a Dirichlet distribution with parameter α. Second, for each word w, a topic z is chosen from this topic distribution. Finally, the word w is generated by randomly sampling from a topic-specific multinomial distribution ϕ_z. The generating probability of word w from document **D** for LDA is given as:

$$P(w|d,\theta,\phi) = \sum_{z} P(w|z,\phi_z)P(z|d,\theta_d) \tag{6}$$

Word Latent Dirichlet Allocation (wLDA). wLDA uses description words posted by all users for a specific resource. In **wLDA**, a resource is viewed as a composition of the description words written by all tagging users. Symbolically, for a resource r we can write it as: $r = \{u_1w_1 + u_2w_2 + \cdots + u_iw_i\}$, where u_iw_i is the description written by a user u_i about a specific resource r. According to basic LDA equation, the generating probability of word w from resources for LDA is given as Eq.7.

$$P_{wLDA}(w|r,\theta,\phi) = \sum_{z} P(w|z,\phi_z)P(z|r,\theta_r) \tag{7}$$

Tags Latent Dirichlet Allocation (tLDA). tLDA uses tags posted by all users for a specific resource. In tLDA a resource is viewed as a collection of the tags given by all users to a specific resource. Symbolically, for a resource r we can write it as: $r = \{u_1t_1 + u_2t_2 + \cdots + u_it_i\}$, where u_it_i is a tag posted by a user about a specific resource. Like as **wLDA**, we calculate the probability for a tag t from the collection by Eq. 8.

$$P_{tLDA}(t|r,\theta,\phi) = \sum_{z} P(t|z,\phi_z)P(z|r,\theta_r) \tag{8}$$

Tag-Topic Model (TT). LDA is extended to model words and authors simultaneously by using latent topics, named Author-Topic Model [13]. We propose its variation in which, each tag of a resource r is associated with a multinomial distribution θ_t over topics is sampled from Dirichlet α, and each topic is associated with a multinomial distribution ϕ_z sampled from Dirichlet β over description words of a resource for that topic. The generating probability of word w for tag t of a resource r is given in Eq. 9.

$$P(w|t,r,\phi,\theta) = \sum_{z} P(w|z,\phi_z)P(z|t,\theta_t) \tag{9}$$

In **TT** model, both words and tags of a resource are modeled together to get their combined influence. It assumes that description words can be real representatives of tags. In its generative process, it considers that words of a resource are responsible for generating some topics which in turn generates the tags for that resource.

User-Tag-Topic Model (UTT). Recently, Author-Topic model is extended to model authors and conferences simultaneously by using latent topics, named Author-Conference-Topic Model (ACT) [14]. Its variation is provided by [8], in which each description word is represented by the probability distribution description over topics, and each topic is represented as a probability distribution over tags and users for each word of a resource for that topic. The generative probability of the tag t with user u for word w of resource r is given as:

$$P(w, u|t, r, \phi, \psi, \theta) = \sum_{z} P(w|z, \phi_z) P(u|z, \psi_z) P(z|t, \theta_t) \tag{10}$$

In UTT, words, tags and users are modeled together to get their combined influence, with assumption that all three components are interdependent and should be modeled together. In its generative process, it considers that words of a resource are responsible for generating some topics which in turn generates the tags and user for that resource.

3.3 Hybrid Approaches

We propose two different approaches to recommend tags: keyword based approach (KBA) and topic modeling based approach (TMBA). These two approaches focus on two different aspects. KBA will recommend some keywords extracted from the candidate set while TMBA tends to recommend some conceptual knowledge as tags. In detail, LM can only model word specific information and cannot model the topic information. LDA models can be used to model the topic distributions of documents, which eliminates the limitation of LM. Consequently, the combination of LM and LDA is an necessary and intuitive better way to improve tag recommendation performance.

Thus, in our hybrid method, we will select and recommend top k ranked tags from a probability list. We define the probability of a tag for the specific resource as an aggregation of KBA and TMBA:

$$\eta(r, t) = \gamma(\delta(r, t), \vartheta(r, t)) \tag{11}$$

where $\delta(r, t) \in [0,1]$ is the probability of a tag for the specific resource by KBA and $\vartheta(r, t) \in [0,1]$ is the probability of a tag for the specific resource by TMBA. γ is an aggregation function used to combing the effects of KBA and TMBA, which should observe the following axioms.

Axiom 1 *For a tag* t *and a resource* r,

- *if* $\delta(\mathrm{r,t})=0$, $\eta(\mathrm{r,t}) = \vartheta(\mathrm{r,t})$.
- *if* $\vartheta(\mathrm{r,t})=0$, $\eta(\mathrm{r,t}) = \delta(\mathrm{r,t})$.

Axiom 2 *For a tag* t *and a resource* r, *function* γ *should guarantee* $\eta(\mathrm{r,t}) \in [0,1]$.

Axiom 3 *For a resource* r, *two tags* t_1 *and* t_2.

- *if* $\delta(\mathrm{r},t_1)=\delta(\mathrm{r},t_2)$ *and* $\vartheta(\mathrm{r},t_1) \geq \vartheta(\mathrm{r},t_2)$, *then* $\eta(\mathrm{r},t_1) \geq \eta(\mathrm{r},t_2)$.
- *if* $\delta(\mathrm{r},t_1) \geq \delta(\mathrm{r},t_2)$ *and* $\vartheta(\mathrm{r},t_1)=\vartheta(\mathrm{r},t_2)$, *then* $\eta(\mathrm{r},t_1) \geq \eta(\mathrm{r},t_2)$.

Axiom 1 specifies the boundary cases of the probability. When $\delta(r,t)=0$, it means that a tag t is in the training file but not in the candidate set of KBA. Thus, $\eta(r,t)$ should be equal to $\vartheta(r,t)$. When $\vartheta(r,t)=0$, it means that a tag t is in the candidate of KBA but not in the training file. Thus, $\eta(r,t)$ should be equal to $\delta(r,t)$. Axiom 2 indicates aggregation function γ should constrain $\eta(r,t) \in [0,1]$. Axiom 3 shows the interrelationship between KBA and TMBA. The aggregation γ is application-dependent [21]. The choice of aggregation function γ in a specific application is out of the scope of this paper[5]. In our paper, we present a possible function as an example to aggregate KBA and TMBA.

$$\eta(r,t) = \xi\delta(r,t) + (1-\xi)\vartheta(r,t) \tag{12}$$

where ξ is a weight to balance KBA and TMBA.

Based on these foundations mentioned-above, we propose several hybrid approaches for tag recommendation on the basis of KBA and TMBA. We instance the KBA as LM and TMBA as LDA, TT and UTT, respectively. Therefore, We consider several forms of combination including **wLM** + **wLDA** (Eq.13), **wLM** + **tLDA** (Eq.14), **wLM** + **TT** (Eq.15), **wLM** + **UTT** (Eq.16), **tLM** + **wLDA** (Eq.21), **tLM** + **tLDA** (Eq.18), **tLM** + **TT** (Eq.19) and **tLM** + **UTT** (Eq.20).

$$P(t|r) = \xi P_{wLM}(t|r) + (1-\xi)P_{wLDA}(t|r) \tag{13}$$

$$P(t|r) = \xi P_{wLM}(t|r) + (1-\xi)P_{tLDA}(t|r) \tag{14}$$

$$P(t|r) = \xi P_{wLM}(t|r) + (1-\xi)P(w|t,r,\phi,\theta) \tag{15}$$

$$P(t|r) = \xi P_{wLM}(t|r) + (1-\xi)P(w,u|t,r,\phi,\psi,\theta) \tag{16}$$

$$P(t|r) = \xi P_{tLM}(t|r) + (1-\xi)P_{wLDA}(t|r) \tag{17}$$

$$P(t|r) = \xi P_{tLM}(t|r) + (1-\xi)P_{tLDA}(t|r) \tag{18}$$

$$P(t|r) = \xi P_{tLM}(t|r) + (1-\xi)P(w|t,r,\phi,\theta) \tag{19}$$

$$P(t|r) = \xi P_{tLM}(t|r) + (1-\xi)P(w,u|t,r,\phi,\psi,\theta) \tag{20}$$

where $P_{wLM}(t|r)$ and $P_{tLM}(t|r)$ is the generating probability of word w and tag t from resource r by language model (as defined in Eq.3 and Eq.4), $P_{wLDA}(t|r)$ and $P_{tLDA}(t|r)$ is the probability of word w or tag t from resource r obtained by the LDA model (as defined in Eq.7 and Eq.8), $P(w|t,r,\phi,\theta)$ is the generating probability of word w for tag t of a resource r, and $P(w,u|t,r,\phi,\psi,\theta)$ is the probability of the tag t with user u for word w of resource r.

4 Experiments

4.1 Experimental Settings

Data sets. Bibsonomy is a social tagging tool which allows user to manage a personal collection of links to the websites and describe those links with one or more words. We selected a subset of Bibsonomy dataset provided by ECML/PKDD 2008 organizers for Discovery Challenge. There are 33,256 words, 13,276 tags, 1,185 users, 14,443 resources and 262,445 bookmarks in total. We then preprocessed dataset by a) removing stop-words, punctuations and numbers, b) lower-casing the obtained words, tags, user names and c) removing words and tags that appear less than three times in the corpus. This creates 18,450 words, 10,702 tags, 861 users, 13,179 resources and 181,491 bookmarks in the dataset.

[5] Please refer to [21] for more detail.

Baseline Approaches. We conducted experiments on Bibsonomy dataset to evaluate the recommendation quality of proposed hybrid and non-hybrid approaches using top-k recommendations metric. We use **wLM**, **tLM**, **CombLM**, **wLDA**, **tLDA**, **TT**, and **UTT** as the baseline methods.

Parameter settings. For keyword based methods we used $\lambda = 10$, as using different values does not have any impact on the performance. For topic modeling methods, one can estimate the optimal values of hyper-parameters α, β and λ by using Expectation Maximization (EM) method [22] or Gibbs sampling algorithm [23]. EM algorithm is susceptible to local maxima and computationally inefficient [12]. Consequently, Gibbs sampling algorithm is fit to use. In our experiments, for 100 topics z the hyper-parameters α, β and λ were set at 50/z, 0.01 and 0.1, respectively, by following values used in [14]. The number of topics z was fixed at 100 on the basis of human judgment of meaningful topics plus measured perplexity [24,25] on held out test dataset for different number of topics z from 20 to 200. For combination, we set $\xi = 0.15$.

Evaluation Measure. Our ultimate goal is to measure the effectiveness of suggesting top-ranked tags for a resource to a user. Thus, to fairly compare their performance, we employ the top-*k* recommendations performance measure. That is, each ranking algorithm needs to recommend the top *k* tags to a resource. This evaluation method is used for community recommendation using LDA by Chen et al [26], in which they ranked randomly withhold communities and recommended top *k* communities.

4.2 Results and Discussions

Topic Related Tags for Resources. We recommend tags with respect to specific kind of resources on the basis of topic layer. Table 1 illustrates 4 different topic related tags out of 100 by LM combined with LDA, discovered from the 1000th iteration of the particular Gibbs sampler run. Each topic shows top 10 tags that are most likely to be produced for topic related resources. The tags associated with each topic are quite intuitive and precise in the sense of conveying a semantic summary of a specific kind of resources. One can see that top ranked tags for different topics are quite informative. For example, in case of topic *Internet Security*, all tags are semantically related and show that resources will have some information about internet security issues. Similarly topics *Music*, *Semantic Web* and *Browser Extension* have tags that are real representatives of specific kind of resources. Our proposed approach discovers several other topics related to *folksonomy*, *graphics*, *data mining*, *web services*, *web tools*, *java*, *book marking*, *business*, *jobs* and also other topics that span the full range of areas encompassed in the dataset. The discovered tag results can be used to suggest meaningful tags for resources and to solve cold start problem by predicting tags, when a user wants to tag new resources to speed up the tagging process by suggesting several tags for each resource.

Top-k Recommendations. In Table 2, we show the performance of proposed approaches and baseline methods, which includes words based, tags based and combined methods. We can see that LM approaches perform better than others, the hybrid approaches is better than non-hybrid approaches except LM. The best average ranking

Table 1. An illustration of top 10 recommended tags for 4 topics(out of 100). Each tag is shown with its probability conditioned on that topic. Topic titles are assigned manually by us to represent the class of tags.

Topic ”Internet Security”	Topic ”Music”	Topic ”Semantic Web”	Topic ”Browsers Extension”
security 0.104911	music 0.139098	semanticweb 0.225512	firefox 0.187719
ittechnology 0.041783	audio 0.093745	rdf 0.103874	extension 0.097383
privacy 0.037503	mp3 0.062511	ontology 0.099498	browser 0.055331
hack 0.020918	media 0.040690	semantic 0.063911	mozilla 0.037030
firewall 0.017173	podcast 0.039406	semweb 0.039992	extensions 0.034304
hacking 0.011823	radio 0.029565	owl 0.034450	bookmarks 0.028853
proxy 0.010753	ipod 0.018013	ontologies 0.020740	greasemonkey 0.024570
safetysecurityprivacy 0.007543	podcasting 0.015018	semantics 0.016948	firefoxextension 0.016393
antivirus 0.007543	itunes 0.013307	sparql 0.014322	best 0.014446
password 0.007008	player 0.010739	semanticmarkup 0.006447	plugin 0.012110

accuracy (k = 2,4,6,8, and 10) of hybrid approach, wLM + tLDA is 65% that is 5% less than the best LM approach, wLM, 70%. We can see that better results can be obtained when keyword and topic layer based approaches are combined to build hybrid approaches. Hybrid approaches having fewer entities, such as wLM + tLDA, tLM + tLDA, perform better, while hybrid approaches modeling more entities, e.g. tLM + TT and tLM + UTT, perform worse than others. It shows that the performance can be increased when keyword and topic layer based methods for words or tags are combined, though LM as individual is still the best. It is assumed that words of resources generate tags of resources in TT approach, and words of resources simultaneously generate tags and users who tagged the resources in UTT approach. We can see that wLDA (55%) and tLDA (57%) perform better than both TT (52%) and UTT (47%), and TT performs better than UTT. It also shows that additional information used for tag recommendation results in poor performance for topic layer based approach, such as users, words and tags are modeled together in case of UTT, and words and tags are modeled together in case of TT. When tLM is combined with wLM, only 60% accuracy is obtained which is lower than that of the original accuracy obtained by tLM (68%) and wLM (70%). That means the hybridization is not useful when both approaches are keyword based ones in the hybrid.

Correlation between Resources. Hybrid approaches can be used for automatic correlation discovery between resources, including keyword and topic layer based approaches influence, simultaneously. Discovered correlations can be utilized to recommend resources to the users with similar interests. To illustrate how it can be used in this respect, the distance between resources i and j is defined as symmetric KL (sKL) divergence [27] between the topic distributions conditioned on each of the resource distribution as:

$$sKL(i, j) = \sum_{z} [\theta_{iz} \log \frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} \log \frac{\theta_{jz}}{\theta_{iz}}] \qquad (21)$$

where θ_{iz} and θ_{jz} are the probability of topic z in resource i and resource j, respectively.

Table 2. Top-k recommendations performance in terms of accuracy by LM plus LDA

Approach	2	4	6	8	10	Average Accuracy
wLM	0.8249148	0.7649713	0.675038	0.633667	0.602264	0.700171
tLM	0.8550615	0.7373653	0.6684583	0.577326	0.600002	0.687643
wLM + tLDA	0.87460833	0.5661024	0.669767	0.6161768	0.567611	0.658853
tLM + tLDA	0.864880	0.5687551	0.669856	0.6127162	0.567574	0.656756
wLM + wLDA	0.8588170	0.7317167	0.657802	0.4472075	0.515749	0.642259
wLM + TT	0.7693678	0.6126987	0.622016	0.6695815	0.480619	0.630857
tLM + TT	0.7622995	0.6210024	0.61584	0.6688483	0.482262	0.63005
tLM + wLDA	0.8074751	0.7160438	0.638026	0.4335037	0.503258	0.619661
CombLM	0.740045	0.6575767	0.5449767	0.55638	0.537967	0.607389
wLM + UTT	0.7862674	0.61301507	0.575937	0.5876752	0.450426	0.602664
tLDA	0.7923222	0.4158538	0.610341	0.5590527	0.519596	0.579433
tLM + UTT	0.7586219	0.5275542	0.557271	0.5871157	0.434969	0.573104
wLDA	0.7068484	0.6338285	0.6252108	0.5582102	0.460180	0.556856
TT	0.50319819	0.4833695	0.558752	0.6289924	0.437467	0.522356
UTT	0.5102204	0.4654382	0.4878531	0.5367335	0.369267	0.473902

Table 3. Top 10 sKL based related resources with Resource ”Macintosh”

Resource ”Macintosh”	Description Words	Tags
http://desktop.google.com/mac	Google Desktop Download for Mac	mac 07system applications bar interface
http://www.google.com/mac.html	Google Pack	google mac software freeware
http://www.apple.com/downloads-/macosx/automator/	Apple - Downloads - Mac OS X - Automator Actions	apple mac macintosh news technology techapplemac
http://osx.iusethis.com/	iusethis mac software: New Releases	mac aggregator apple
http://www.mecheng.adelaide.edu.-au/ will/texstart/	What do I need for TeX on Mac OS X?	apple latex mac osx tex
http://www.opensourcemac.org/	Open-Source software for OS X applications directory freeware	mac opensource osx
http://www.delicious-monster.com	Delicious Library	digital apple tool software library macosx
http://www.wilber-loves-apple.org/	Wilber loves Apple.We bring the Gimp to the Mac	dmg download gimp macos
http://www.digital-web.com/articles/mac-screencast-capturing/	Digital Web Magazine Capture a Screen cast with a Mac how to capture a screen cast on mac	mac screencast
http://firefoxmac.furbism.com/	Firefox: Mac PPC Optimized Builds	apple browser firefox mac mozilla
http://bibdesk.sourceforge.net/	BibDesk JabRef for Mac OS X	mac bibtex osx

We calculated the dissimilarity between the resources by using Eq.21. The smaller dissimilarity value means the higher correlation between resources. Table 3 shows semantic-based correlation of resource *Macintosh* which is Google desktop download for Mac and tagged by *mac 07 system applications bar interface*. Here, it is obligatory to mention that top 10 resources related to resource *Macintosh* are not the necessary resources which are tagged by similar users, but rather the resources that tend to produce most tags or words for the same topics. Again the results are quite promising and realistic as most of the resources related to resource *Macintosh* are also topic related.

5 Conclusions

This study deals with the problem of tag recommendation using several hybrid and non-hybrid approaches. The approaches are proposed by combining keyword based (no-semantics) and topic layer (semantics) based approaches. In our experiments the performance difference between words and tags based approaches, e.g. tLDA and wLDA, is not very different. Although tLDA is a bit better than wLDA but we can say that both description words and tags can be a good choice for tagging resources. We find that combination of keyword and topic layer based approaches is significantly effective in many cases when there are fewer entities are modeled in topic model. The proposed approaches are simple, but the experiments show that they produced quite intuitive results. We also show that the performance decreases if the number of entities (e.g. words, tags and users) is simultaneously modeled in topic layer based approach. Moreover, through using our approaches, the correlations between the resources can be identified quite precisely.

Acknowledgements. This work is supported by NSFC (60873225, 60773191, 70771043), 863 Program (2007AA01Z403), NSF of Hubei (2009CDB298), Wuhan Chenguang Program (200950431171), Open Foundation of SKLSE (SKLSE20080718), Innovation Fund of HUST (Q2009021), Youth Fund of HBUE (XJ2009013), Humanities and Social Sciences Program of Hubei (2010Q094), Hong Kong RGC (CityU 117608).

References

1. Lipczak, M.: Tag recommendation for folksonomies oriented towards individual users. In: ECML PKDD Discovery Challenge, pp. 84–95 (2008)
2. Hsu, M., Chen, H.: A method to predict social annotations. In: CIKM, pp. 1375–1376 (2008)
3. Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 506–514. Springer, Heidelberg (2007)
4. Mishne, G.: Autotag: a collaborative approach to automated tag assignment for weblog posts. In: WWW, pp. 953–954 (2006)
5. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Folkrank: A ranking algorithm for folksonomies. In: LWA, pp. 111–114 (2006)
6. Sood, S., Owsley, S., Hammond, K., Birnbaum, L.: Tagassist: Automatic tag suggestion for blog posts. In: ICWSM (2007)
7. Tatu, M., Srikanth, M., DSilva, T.: Tag recommendations using bookmark content. In: RSDC, pp. 96–107 (2008)

8. Zhang, N., Zhang, Y., Tang, J.: A tag recommendation system based on contents. In: RSDC (2008)
9. Chen, Z., Cao, J., Song, Y., Guo, J., Zhang, Y., Li, J.: Context-oriented web video tag recommendation. In: WWW (to appear, 2010)
10. Daud, A., Li, J., Zhou, L., Muhammad, F.: A generalized topic modeling approach for maven search. In: Li, Q., Feng, L., Pei, J., Wang, S.X., Zhou, X., Zhu, Q.-M. (eds.) APWeb/WAIM 2009. LNCS, vol. 5446, pp. 138–149. Springer, Heidelberg (2009)
11. Wei, X., Croft, W.B.: Lda-based document models for ad-hoc retrieval. In: SIGIR, pp. 178–185 (2006)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: NIPS, pp. 601–608 (2001)
13. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.L.: Probabilistic author-topic models for information discovery. In: KDD, pp. 306–315 (2004)
14. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: KDD, pp. 990–998 (2008)
15. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: WWW, pp. 501–510 (2007)
16. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: WWW, pp. 211–220 (2007)
17. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW, pp. 327–336 (2008)
18. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Mining association rules in folksonomies. In: IFCS, pp. 261–270 (2006)
19. Ruch, P., Baud, R.H., Geissbühler, A.: Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. I. J. Medical Informatics 67(1-3), 75–83 (2002)
20. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: SIGIR, pp. 334–342 (2001)
21. Lesot, M.J., Mouillet, L., Bouchon-Meunier, B.: Fuzzy prototypes based on typicality degrees. In: Fuzzy Days, pp. 125–138 (2004)
22. Hofmann, T.: Probabilistic latent semantic analysis. In: UAI, pp. 289–296 (1999)
23. Griffiths, T., Steyvers, M.: Finding scientific topics. PNAS 101, 5228–5235 (2004)
24. Azzopardi, L., Girolami, M., van Rijsbergen, K.: Investigating the relationship between language model perplexity and ir precision-recall measures. In: SIGIR, pp. 369–370 (2003)
25. Daud, A., Li, J., Zhou, L., Muhammad, F.: Knowledge discovery through parametric directed probabilistic topic models. Journal of Frontiers of Computer Science in China (2009)
26. Chen, W., Chu, J.C., Luan, J., Bai, H., Wang, Y., Chang, E.Y.: Collaborative filtering for orkut communities: discovery of user latent behavior. In: WWW, pp. 681–690 (2009)
27. Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI, pp. 487–494 (2004)
28. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of IR techniques. ACM TOIS 4(20), 422–446 (2002)
29. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: KDD, pp. 426–434 (2008)
30. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision-making. IEEE Trans. Syst. Man Cybern. 18(1), 183–190 (1988)