Modeling Ontology of Folksonomy with Latent Semantics of Tags

Ali Daud
Department of CS & Technology
1-308, FIT Building, Tsinghua University
Beijing, 100084, China
ali_msdb@hotmail.com

Juanzi Li
Department of CS and Technology
10-206, East Main Building, Tsinghua University
Beijing, 100084, China
ljz@keg.cs.tsinghua.edu.cn

Lizhu Zhou
Department of CS and Technology
3-409, FIT Building, Tsinghua University
Beijing, 100084, China
dcszlz@tsinghua.edu.cn

Lei Zhang
Graduate School at Shenzhen, Tsinghua University
F-303 B, Tsinghua Campus
Shenzhen, 518055, China
zhanglei@sz.tsinghua.edu.cn

Ying Ding
School of Library and Information Science, Indiana University
1320 E 10th
Bloomington, USA
dingying@indiana.edu

Faqir Muhammad
Department of Mathematics & Statistics,
Allama Iqbal Open University, Sector H/8
Islamabad, 44000, Pakistan
aioufsd@yahoo.com

Abstract

Modeling ontology of folksonomy provides a way of learning light weight ontology's which is a hot topic investigated recently. Previous approaches for modeling ontology of folksonomy either ignores semantics (synonymy, hyponymy or polysemy) or do not simultaneously consider relationships between actors (users), concepts (tags) and instances (resources) or are based on the idea that title words are responsible for generating tags for resources. Latent semantics and user-tag dependencies instead of user-word dependencies however are extremely important. In this paper we address these problems by introducing a latent topic layer into the traditional tripartite Actor-Concept-Instance graph. We thus propose an Actor-Concept-Instance-Topic (ACIT) approach to model ontology from folksonomy in a unified way by directly using tags and users of resources. We illustrate on Bibsonomy dataset that our proposed approach ACIT outperforms title words based approaches Tag-Topic (TT) and (User-Word-Topic) UWT for modeling the ontology of folksonomy.

Keywords
Latent Semantics, User-tag Dependencies, Light Weight Ontology's, Folksonomy, Unsupervised Learning

1. Introduction

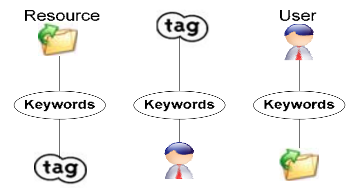
Social tagging systems allow users to store and share various types of resources on the internet in the systems such as Flicker, YouTube, Bibsonomy and Delicious. One of the major outputs of this user-tag-resource activity is called a folksonomy. Different resources are tagged with a variety of tags by different users. Intuitively, similar tags and the users tagging similar resources both can be used to create a bridge between resources. Folksonomies have therefore become so called "user-generated ontologies" in the Semantic Web understanding.

Notably, these tags are free will keywords or uncontrolled vocabularies added by Web users; where synonymy (multiple tags expressing the same meaning, e.g., "Data Mining" and "Knowledge Discovery"), homonymy (a single tag "party" used with different meanings e.g. political party and farewell party) and polysemy (a single tag used with multiple related meanings) are common. Search is an important example which explains the importance of synonymy and homonymy and user-word dependencies for learning light weight ontology's. Search results are usually restricted to the specific tags used in the process of annotation, while linguistic and semantic limitations of tags affect the search capabilities. e.g., if a user assigns a tag "dog" to a resource, and another one looks for the word "animal", that resource will not be shown. Also there are implicit relationships between entities which influence semantic structure of folksonomies, e.g. a user related to "news" word can belong to two different topics or areas such as news technology which spotlight the technologies used in the news and regional news which spotlight the news about different regions. Therefore user-word dependencies with respect to users' interests are considered necessary to deal with the correct use of word "news". These uncontrolled vocabularies trigger problems of reliability, consistency and relationship dependencies in modeling ontology of folksonomy, which must be considered for various applications in the folksonomies such as search, annotation and recommendation tasks.

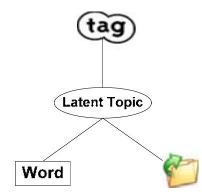
Approaches used to date for modeling the semantic structure of folksonomies can be divided into two major types (1) *independently modeling ontology of folksonomy without actors influence (*The approaches which do not utilize the actors (users) information when modeling ontology of folksonomy*)* with recent investigative effort [27] and (2) d*ependently modeling ontology of folksonomy with actors' influence* with recent investigative effort [20], where both kind of approaches use either keywords or latent topic layer. Yet they either ignore latent semantics or do not simultaneously consider relationships between all social dimensions which are actors, concepts and instances. In the real world, however, users with similar interests usually assign similar tags to annotate similar resources [5,15] where natural relationships existing between them should thus be modeled simultaneously.

In this paper we address these problems by introducing a latent topic layer into a tripartite Actor-Concept-Instance graph [26] to simultaneously capture synonymy, homonymy and relationships between social dimensions. We propose a latent topic layer based Actor-Concept-Instance-Topic (ACIT) approach to d*ependently model ontology from folksonomy* in a unified way as shown in Figure 1(c)]. Figure 1(a) shows recently proposed keywords-based *dependently modeling ontology of folksonomy* (Actor-Concept-Instance (ACI)) approach [20]. In ACI to its limitations only single keyword is used as a bridge and tripartite graph is divided into three bi-partite graphs, therefore ternary relationships between social dimensions are not captured. Figure 1 (b) shows latent topic layer based *independently modeling ontology of folksonomy* (Tag-Topic (TT)) approach [27]. In TT approach latent topic layer is used but to its limitation relationships between all social dimensions are not modeled simultaneously.

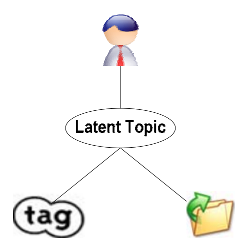
In our proposed ACIT approach latent topic layer (only top ten tags are shown here) is used in addition to simultaneously modeling all social dimensions, which can be more useful to deal with the problems of synonymy, hyponymy and polysemy by using other tags in the same topic and user-word dependencies. Our folksonomy modeling shows that ACIT performed much better than the baseline approaches in terms of accuracy for predicting ranks for existing ontology of folksonomy. Our approach is quite general and requires no specific domain knowledge so can be applied to many different domains. It can also be used for learning the hierarchical semantic structure of folksonomies by combining it with the method proposed in [27].



(a) Keywords-based *dependently modeling the ontology of folksonomy* **(ACI).**



(b) Latent Topic Layer based *independently modeling the ontology of folksonomy* **(TT)**.



Latent Topic Layer	
Tags	***Prob.***
Natural	**0.121514**
Language	**0.059027**
Parsing	**0.0547**
Word	**0.052969**
Processing	**0.050199**
Dependency	**0.045179**
Information	**0.03237**
Grammar	**0.025447**
Tagging	**0.020773**
Sense	**0.020081**

c) Latent topic layer based *dependently modeling the ontology of folksonomy* **(ACIT).**

Figure 1. Visual comparison of most up-to-date previous and proposed approaches.

The contributions of our work described in this paper are the followings:

(1) Mixed the basic idea of ACI model to consider all social dimensions (actors, concepts and instances) and with the basic idea of TT approach to use latent topic layer.

(2) Considered all social dimensions simultaneously, to avoid reducing tri-partite graph to bi-partite graph, which facilitated us to successfully model ternary relationships.

To the best of our knowledge, we are the first to deal with modeling the ontology of folksonomy problem by proposing unified topic modeling approach with directly using tags instead of title words.

The rest of the paper is organized as follows. Section 2 illustrates our proposed approach to model the semantic structure of Folksonomies. Section 3 discusses dataset, parameter settings, evaluation method, and modeled ontology from Folksonomies. Section 4 provides related work and section 5 concludes this paper.

2. Modeling Ontology of Folksonomy

2.1. Folksonomy Graph

In order to learn networks of folksonomies at a semantic level, we represent a tripartite graph with links, where these links are obtained by using the latent topic layer. The set of vertices is partitioned into three (possibly empty) disjoint sets of users as actors A = $\{a_1,a_2,\ldots,a_k\}$, tags as concepts C = $\{c_1,c_2,\ldots,c_m\}$ and resources (objects) as instances I = $\{s_1,s_2,\ldots,s_l\}$ with an additional latent topic layer $Z = \{z_1,z_2,\ldots,z_j\}$ to capture semantic relationships.

In fact, we extend traditional tripartite model [26] of social networks and semantics (actors, concepts and instances) by introducing a latent topic layer. In social tagging systems, users tag objects with concepts that creates a ternary relationship between the actors, concepts and instances. Thus the ontology from folksonomy can thus be defined as a set of annotations $F \subseteq A \times C \times I$, with a latent topic layer as a connecting bridge. By generating concepts for similar resources, the actor's association with that resource and other actors who behave in a similar way are revealed. Using a latent topic layer-based Actor-Concept-Instance-Topic approach (ACIT), we are able to model the relationships between actors and concepts (AC), concepts and instances (CI) and instances and actors (IA).

2.2. Actor-Concept-Instance-Topic (ACIT) Approach

Before explaining our proposed ACIT approach in detail, it is useful to briefly introduce Latent Dirichlet Allocation.

The fundamental topic modeling approach Latent Dirichlet Allocation (LDA) [4] assumes that there is a hidden topic layer $Z = \{z_1, z_2, z_3, \ldots, z_t\}$ between the word tokens and documents, where z_i denotes a latent topic and each document d is a vector of N_d words $\mathbf{w}_d$ with documents vocabulary of size V. 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:

$$P(w|d,\theta,\emptyset) = \sum_{z=1}^{T} P(w|z,\emptyset_z)P(z|d,\theta_d) \qquad (1)$$

The basic idea presented in the Author-Topic model [22] an extension of LDA with adding author dimension, that words and authors of documents can be modeled by considering latent topics became the intuition of modeling actors, concepts and instances in folksonomies, simultaneously. Our intuition is based on the fact that the co-authors of a research paper have the same research interests; intuitively, the users tagging the same kind of resources have similar interests too. For example, a person interested in sports will tag sports websites and a person interested in songs will tag music websites. One the basis of provided intuition; we propose ACIT approach, in which an instance is a composition of its concepts given by all actors. Symbolically, for an instance I we can write it as: $I = \{(\mathbf{c}_1,\mathbf{a}_{L1}) + (\mathbf{c}_2,\mathbf{a}_{L2}) + (\mathbf{c}_3,\mathbf{a}_{L3}) + \ldots + (\mathbf{c}_l,\mathbf{a}_{Lk})\}$, where c_L are concepts of an instance and a_{Lk} are actors for concepts c_l.

The proposed approach follows the natural order of conceptual thought considering that an actor is responsible for generating some latent topics of the instances on the basis of semantics-based information present in the concepts as well as co-instance (tagging the same resource) based associations. In ACIT, each actor (from set of A actors) of an instance is associated with a multinomial distribution θ_r over topics and each topic is associated with a multinomial distribution Φ_z over concepts of a resource for that topic. Both θ_r and Φ_z have symmetric Dirichlet prior to hyper parameters α and β. The generating probability of the concept c for actor r of an instance s is given as:

$$P(c|r,s,\emptyset,\theta) = \sum_{z=1}^{T} P(c|z,\emptyset_z)P(z|r,\theta_r) \qquad (2)$$

$$P(z_i = j, r_i = k | c_{\boldsymbol{i}} = m, \mathbf{z}_{-i}, \boldsymbol{r}_{-\boldsymbol{i}}, \mathbf{a}_{\boldsymbol{s}}) \propto \frac{n_{-i,j}^{(ci)}+\beta}{n_{-i,j}^{(.)}+C\beta} \frac{n_{-i,j}^{(ri)}+\alpha}{n_{-i,.}^{(ri)}+A\alpha} \qquad (3)$$

Gibbs sampling is used [1] for parameter estimation in ACIT which has two latent variables z and r. The conditional posterior distribution for z and r is given by using Eq. 3, where $z_i = j$ and $r_i = k$ represent the assignments of the i^{th} concept of an instance to a topic j and actor k respectively, $c_i = m$ represents the observation that i^{th} concept is the m^{th} concept in the lexicon and z_{-i} and r_{-i} represents all topic and actor assignments not including the i^{th} concept. Furthermore, $n_{-i,j}^{(ci)}$ is the total number of concepts associated with topic j, excluding the current instance, and $n_{-i,j}^{(ri)}$ is the number of times actor k is assigned to topic j, excluding the current instance, where C is the size of the lexicon and A is the number of actors. "." Indicates summing over the column where it occurs and $n_{-i,j}^{(.)}$ stands for number of all concepts that are assigned to topic z excluding the current instance.

During parameter estimation the algorithm only needs to keep track of C x Z (concept by topic) and Z x A (topic by actors) count matrices. From these count matrices topic-concept distribution Φ and actor-topic distribution θ can be calculated as:

$$\emptyset_{zc} = \frac{n_{-i,j}^{(ci)} + \beta}{n_{-i,j}^{(.)} + C\beta} \tag{4}$$

$$\theta_{rz} = \frac{n_{-i,j}^{(ri)} + \alpha}{n_{-i,.}^{(ri)} + A\alpha} \tag{5}$$

where, $\emptyset_{zc}$ is the probability of concept c in topic z and θ_{rz} is the probability of topic z for actor r. To find Z x I (topic by instance) count matrix we calculated the distribution of topic given instance as:

$$P(z|s) = \sum_{r \epsilon A_s} P(z|r) P(r|s) = \frac{1}{|A_s|} \sum_{r \epsilon A_s} P(z|r) \tag{6}$$

where, r_s is the number of actors belonging to an instance s.

2.3. Semantics and User-Dependency Abilities

The latent topic layer based approach that models all entities together can be very useful for modeling ontology of folksonomy. For example, one can see that in Table 1 news and media are (synonymic) tags, both assigned to the "News Technology" topic as they have similar meaning in this context. But these words are not both present in the "Regional News" topic related only to news for different regions. Secondly, one can see that in Table 1, tag "news" is present in two different topics "News Technology" and "Regional News. The word "News" therefore has a different usage (homonym) in both topics; the word is thus used by at least two different kinds of users based on their interests, where all users for both topics are also different. This demonstrates how the immediate modeling of users and resource tags is very important for capturing synonymy or homonymy.

Table 1. Semantics and user-dependency.

"News Technology"				"Regional News"			
Tags	**Prob.**	**User Id**	**Prob.**	**Tags**	**Prob.**	**User Id**	**Prob.**
news	0.34	1747	0.908	news	0.13	246	0.755
technology	0.13	1345	0.078	politics	0.08	697	0.115
tech	0.08	2203	0.005	german	0.06	2932	0.097
daily	0.05	75	0.002	economics	0.05	1249	0.006
magazine	0.03	862	0.001	business	0.04	1976	0.004
media	0.02	1951	0.000	nachrichten	0.04	637	0.002
firefoxrss	0.02	283	0.000	politik	0.03	2069	0.002
it	0.01	1072	0.000	germany	0.03	231	0.001
geek	0.01	438	0.000	finance	0.03	2323	0.001
gadgets	0.01	421	0.000	international	0.03	623	0.000

In Table 1 the users' probabilities are much skewed, by analyzing the dataset we have found that these users have tagged enormously the topic specific related resources. These users can be spammers but we are not focused on spam detection issue here.

3. Experiments

3.1. Experimental Settings

3.1.1. Dataset. Bibsonomy is an online social tagging system. We used bibsonomy dataset herein provided by the ECML/PKDD 2008 organizers for Discovery Challenge. There are 33256 words, 13276 tags, 1185 users, 14443 resources and 41268 bookmarks in total. We then preprocessed dataset by (a) removing stop-words and punctuations (b) lower-casing the obtained words and (c) removing tags, words and users that appear less than three times in the corpus. This led to 6215 tags, 3285 words, 728 users, 13734 resources in the dataset.

3.1.2. Parameter Settings. The optimal values of hyper-parameters α and β for ACIT can be estimated by using Expectation-Maximization [14] or Gibbs sampling algorithm [1]. In our 1000 iterations of Gibbs sampling algorithm based experiments, for 80 topics Z the values of hyper-parameters α and β are set at $50/Z$ and 0.01 [11]. The number of topics Z is fixed at 80 on the basis of human judgment of meaningful topics and measured perplexity [2,11] on 20% held out dataset for different number of topics for Z from 2 to 200.

3.1.3. Evaluation Method. It inevitably requires consulting the community or communities whose conceptualizations are being learned, a time consuming task. After identifying the superiority and usefulness of these conceptualizations, we evaluated our proposed approach by showing its accuracy for the ranking prediction of original tags and users of the resource (in other words we can say comparing existing ontology with modeled one). Our ultimate goal is to measure the effectiveness of prediction ranking of objects (tags or users) for a resource. To effectively compare their performance, we thus employ the top-*k* (of size *k*, which is 2,4,6,8 and 10 in this work) predicted ranking performance measure. That is, each prediction ranking algorithm needs to rank the top *k* objects. This evaluation method is adopted from the evaluation method used for community recommendation using Latent Dirichlet Allocation by Chen et al. [9], where they ranked randomly withheld communities.

3.1.4. Baseline Approaches. We compare our proposed ACIT approach with tag-topic (TT) model [27], which is a variation of rather complex topic-tag model [7] that has same idea of using words of resources to generate tags but with more complex structure. And user-word-topic (UWT) approach which is a variation of rather complex user-topic-tag model [7] that has same idea of using words and users of resources simultaneously to generate tags but with more complex structure for the existing ontology and modeled ontology of folksonomy. Here, existing ontology means the tags and users already attached with the resource and modeled ontology means the predicted ranking of the withhold tag and user for that resource. We apply the same number of topics for comparability, where the number of Gibbs sampler iterations and parameter values used for TT and UWT are also the same as those used for ACIT.

The TT approach considers that title words are responsible for generating tags for the resources. Each word (from a set of title words) of a resource is thus associated with a multinomial distribution θ_w over topics is sampled from Dirichlet α, and each topic is associated with a multinomial distribution Φ_z sampled from Dirichlet β over tags of a resource for that topic.

The UWT approach considers that users are responsible for generating tags for resources. Each user (from a set of users) of a resource is thus associated with a multinomial distribution θ_r over topics is sampled from Dirichlet α, and each topic is associated with a multinomial distribution Φ_z sampled from Dirichlet β over words of an instance for that topic. Because users usually tag resources with keywords present in the words attached to them, we presume here that title words of resources are likely to be tags of a resource.

3.2. Results and Discussions

3.2.1. Modeled Ontology of Folksonomy. We used the ACIT approach herein to model latent semantics and relationships between entities in a unified way. Consequently, we have generated as Actor-Concept-Instance graph instead of generating separate graphs of Actor-Concept and Concept-Instance [20]. We used Actor-Concept-Instance graph to model O_{aci} ontology from folksonomy. The ontology O_{aci} is based on actors sharing concepts as interests (i.e. the associations reflect overlapping of tags used by actors) and concepts that share instances as communities (i.e. the associations reflect co-occurrence of tags for similar resources).

We have filtered the network based on the absolute strength of probabilistic associations between entities. Table 2 shows a view of the O_{aci} graph with results giving clear evidence of emerging semantics in the network. We illustrate five different clusters of interests out of eighty, discovered from the 1000th iteration of particular Gibbs sampler run. Analysis reveals that the top objects having high probabilities in clusters are often very specialized terms, while the bottom objects having low probabilities are overly general terms.

Table 2: Five main clusters of interest (top ten for each cluster) based on concept-topic and actor-topic network (Titles are our interpretation of the clusters).

Topic Title	**Concepts**
Data Mining	Statistics, data, **datamining**, mining, clustering, ranking, ml ,dataset, **machinelearning**, ki2007webmining
Semantic Web	**semanticweb**, rdf, ontology, semantic, Owl, foaf, **semweb**, sparql, metadata, ontologies
Web Design	html, **webdesign**, **webdev**, css, cms, xhtml, php, w3c, markup, **webdevelopment**
Music	music, audio, mp3, media, **podcast**, radio, Streaming, ipod, **podcasting**, **music**
Photo	flickr, photos, photography, photo, images, image, landscapes nature, **foto**, photographs
Title	**Actors**
Data Mining	**3**, 438, 1888, 2861, 71, 1746, 1900, 217, 2706, 883
Semantic Web	**524**, 14, 2075, 1888, 49, 293, 787, 1129, 1063, 1703
Web Design	2008, 491, 1015, 1152, 593, 3383, 1192, 564, 647, 152
Music	**2977**, 421, 2128, 155, 105, 586, 1045, 1732, 540, 597
Photo	421, 1045, 2017, 2979, 978, 155, 351, 86, 2488, 61

The process of tagging is made as easy as possible. A textbox allows actors to enter a set of words without any recommendations or restrictions made by the system. Consequently, synonyms are common in the folksonomy, e.g. “semanticweb”, “semweb” and “webdevelopment” and “webdev”are different keywords in Table 2. Ambiguity is also present, since users often pick short terms to describe items, such as “ml” for "machine learning” in Data Mining concepts, where the ml keyword becomes meaningful because of other related keywords, notably machinelearning keyword in the same concept. Furthermore, users often make the mistake of entering key phrases instead of keywords (e.g. “Data Mining”), where the words are subsequently parsed as separate tags ("Data" and "Mining"); or they escape one word limitations by concatenating words e.g. “semanticweb”. Both of these problems are effectively handled by our approach as seen in the concepts “Data Mining” and “Semantic Web”, respectively. Different language, abbreviated or alternative spellings and meanings for keywords are also an issue. For example, one may find “musik” and “foto” in Music and Photo concepts which are words used in German language, or the currently used “dialog” instead of “dialogue,” all of which are correctly associated with related clusters, despite the assumed language differences or varied spellings. As words shift in meaning or popularity, they may also lose connectivity, such as the term “gay,” which is rarely used in the traditional meaning of “happy” due to potentially misleading or unwanted associations.

The concepts associated with each topic are strongly semantically related, illustratively, and keywords associated with “Music” concept and all other concepts discovered by ACIT are very much clear in describing different aspects of music. Consequently, the actors associated with the concepts are also intuitive. For example, by analyzing dataset and results, we find that the top userId 3 for “Data Mining” concept tagged 639 resources from the total of 13734 and is assigned by ACIT to seven different concepts in top ten users for each topic (**data mining, natural language processing, web services,** maps, research meetings, **search engines** and Folksonomies), in which five topics shown in bold font above are somehow related to each other as research areas showing that the user is active in these areas. Additionally, the user utilized maps to arrange his research meetings (e.g. conferences and workshops he/she attended) by applying Folksonomies. We do not know the user name here because of name encoding, but we are certain that the user is a very active person in the aforementioned research areas. Top userId 524 for semantic web topic tagged 167 resources and is assigned to only one topic in top ten users for each topic (semantic web) by ACIT. By analyzing the resources tagged by him we have found that he just tagged resources related to semantic web (shows highly specific behavior). Top userId 2977 for music topic tagged 527 resources and is assigned to 4 different topics in top ten users for each topic (music, media tools, internet security and delicious) which shows that he is a common user and likes to listen music, play with media tools with a bit interest in internet security issues.

Here it is obligatory to mention that top 10 users are not necessarily the most active taggers in that community, but rather are the actors that are semantically related to the topics, which build up topic based community.

In case of keywords-based ACI model [20] by splitting tripartite graph into two bipartite graphs causes failure to simultaneously model ternary associations, which are needed to capture polysemy and homonymy as explained in Table 1 and 2. In case of TT approach [27] the assumption that words are responsible for generating tags has conflict with real situation in which users are generating tags. Finally, since understanding of the ontology of folksonomy is affected by many factors, here the latent semantics and actors influence only means, some *potential* to be important in comparison with previous approaches in this context.

As Gibbs sampling is time consuming so running model for each new resource is computationally expensive. For this purpose only Eq. 3 can be applied on each new resource for temporarily updating the count matrices by using just 10 iterations in our simulations which takes less than 3 seconds.

3.2.2. Accuracy of Modeled Ontology of Folksonomy

We show the effectiveness of our proposed approach ACIT for modeling the ontology of folksonomy in terms of accuracy. ACIT approach performed better as compared to TT and UWT approaches. The average accuracy results for modeled ontology of folksonomy for tag ranking prediction for k= 2,4,6,8,10 and number of topics varied from 20, 40, …,200 shown in Table 3 are .525 for ACIT, .451 for TT and .441 for UWT, which show that ACIT approach performed 7.4% and 8.4% better than TT and UWT approach in terms of accuracy which show the better performance of our proposed approach. The average accuracy results for user ranking prediction are .574 for the ACIT and .449 for the UWT which show that our proposed ACIT approach performed 12.5% better than UWT approach in terms of accuracy which is highly significant. Collectively one can say that modeling users and tags (semantics and user-dependencies) of resources together is useful and effective.

Table 3. Accuracy for the modeled ontology of folksonomy.

Average Accuracy	UWT	TT	ACIT
Tag	0.441	0.451	**0.525**
User	0.449	NA	**0.574**

By analyzing tagging system we have found that title words used in TT (a variation of topic tag model [27]) and UWT (a variation of user-topic-tag model [7])) are rather general while, users usually assign different specific tags to different pages on the same URL. For example, SWFUploadbeta
http://labb.dev.mammon.se/swfupload
URL has SWFUpload beta title words which are very general as compared to tags assigned by user 122 (assigned 5 tags), user 884 (assigned 4 tags) and user 14 (assigned 2 tags) to the different pages of same URL.

122 flash programming upload web20 webdesign
884 programming upload web20 webdesign

14 flash upload

Intuitively, one can see that webdesign, flash upload and programming tags are not very general and are shared at least between two users which support our thinking that modeling tags association by utilizing the user-dependencies is important. This is just a small example with only 3 users tagged this URL. In other situations a URL can be tagged by more than ten users which make the tags more specific and user-dependencies more influential.

3.2.3. Concepts Correlation Analysis

ACIT approach can be used for correlation discovery between concepts, including actors influence in comparison to previously used influence of only words [22,27]. Discovered correlations can be utilized to find synonyms and semantically related concepts. To illustrate how it can be used in this respect, distance between concepts i and j is defined as symmetric KL (sKL) divergence between the latent topics distribution conditioned on each of the concepts distribution as:

$$sKL(i,j) = \sum_{z=1}^{T} \left[\theta_{iz} log \frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} log \frac{\theta_{jz}}{\theta_{iz}}\right] \qquad (7)$$

Table 4. Concepts correlation analysis.

Concept	Related Concepts
Nokia	pda,shareware,calls,telephony,phones,handy,telephone,s60,pocketpc,vergleich, cheap,jahjah,sip,ct,telefon,forsfreitas,exif,phonecalls,mobilfunk,clipart
Graph	analysis,sna,statistics,visualisation,infovis,timeline,statistik,visual,graphs,social networkanalysis,stats,charts,djembe,chart,infographics,djembes,graphviz,navigation,visualisierung,graphtheory
Business	people,enterprise,economics,finance,money,company,advertising,enterprise,companies,startup,government,virtualization,financeeconomicsmanagement,economy,babsonfip,adsense,strategy,unternehmen,group,press,intranet
Language	dictionary,english,naturallanguageprocessing,translation,thesaurus,wörterbuch, deutsch,lexicon,sprache,languages,lexikon,dict,englisch,lingue,dictionaries,linguistics,words,italian,grammar,encyclopedia,synonyme
Security	ittechnology,privacy,lifehacker,safetysecurityprivacy,firewall,monitoring,datenschutz,tipstricks,hacking,proxy,encryption,tracking,admin,anonymous,ip,support,sicherheit,password,router,überwachung,securite

Dissimilarity between the concepts is calculated by using Eq. 11; smaller dissimilarity value means higher associative relationship between the concepts. Table 4 shows latent semantics and user-dependencies based correlations for different concepts, in which all concepts are shown in order (having smallest value at first on left-side and so on). Here, it is obligatory to mention that top 20 concepts shown are not just the concepts that have co-occurred with that concept for similar instance mostly, but also the concepts that tend to be assigned by similar actors to other instances. It is quite obvious, that top 20 related concepts with the concept have similar sense in different respects and covers a domain specific knowledge very well.

For example, for concept “Nokia” found related concepts are; pda is (Nokia PDA phone), shareware (free Nokia software), calls, telephony (Nokia, Intel dial up open source project), phones, handy (Nokia phones property), telephone, s60 (a software platform for mobile phones that runs on Symbian OS), pocketpc (Nokia pocket PC), vergleich (Nokia Handy), cheap, jahjah (a famous ringtone for Nokia), sip (Nokia session initiation protocol is a signaling protocol) and others provides us with a very handy vocabulary of keywords, which are highly domain specific and look to be engineered by members of a domain.

4. Related Work

Social tagging systems have provided Web with rapidly growing social networks. Associations’ growth between users in these networks is exponential, and tags assigned by users are providing us with keywords, so-called uncontrolled vocabularies. A few efforts have been made to automatically model light weight ontology from folksonomies by capturing synonym and homonym relations between tags. Previous efforts used a wide variety of linguistic, rule-based and clustering-based approaches.

An approach for effectively browsing large scale web annotations is proposed [18]. Clustering is performed [26] to make clusters of highly related tags where each cluster is associated with a concept of the existing ontology. A unified model ACI of social networks and semantics is proposed by arguing that we are ontologies in social networks [20]. It extends traditional bi-partite ontology graph to tripartite graph by introducing actors’ social dimension. Intuitively semantics and associations emerge from users (actors) annotating resources in tagging systems are considered important as appropriate. Recently, a tag-topic (TT) approach is proposed to model tags with the help of title or description words of a resource [27]. A topic-tag and user-topic-tag models [7] with more complex structures are introduced based on the similar idea of tag-topic model [27] that title words are useful for generating the tags for resources.

Previously modeling of all social dimensions is utilized in academic social networks for capturing the correlations effect for expert finding problem [10,12]. Latent layer based simultaneous modeling of all social dimensions in social tagging systems is introduced here. Our proposed ACIT approach is capable of modeling latent semantics and dependencies (relationships) between all tagging social network dimensions, simultaneously and proved to be effective in comparison to approaches using title words for generating tags of resource.

With the emergence of social tagging systems several applications has emerged, such as friend recommendation [3,19]. The quality of the metadata and the scalability compared with conventional indexing systems for social tagging systems is discussed [16]. From Indexing and retrieval application it is found that if vocabulary terms used are from authoritative source significant advantages can be obtained [13]. Several algorithms are developed for recommending mood and theme annotations in order to support users in tagging [6]. Social trust importance in online communities is highlighted [8].

Finally we can say that our proposed approach is quite general and realistic, therefore applicable to most of the aforementioned applications in the folksonomies by defining problem setting in an appropriate way.

5. Conclusions

This study shows that modeling ontology of folksonomy with latent semantics by simultaneously dealing with actors, concepts and instances without using title words is significant. Our proposed Actor-Topic-Instance-Topic approach utilizes these factors and proves its effectiveness in the bibsonomy dataset. It is evident that by using latent semantics and modeling dependencies between all social dimensions one can get more precise ranking results of modeled ontology of folksonomy. Additionally, the demonstrated associative relationships between concepts are precise and functional.

6. Acknowledgements

The work is supported by the National Natural Science Foundation of China under Grant (60973102, 60703059) and Chinese National Key Foundation Research and Development Plan under Grant (2007CB310803).

7. References

[1] Andrieu, C., Freitas, N.D., Doucet, A., and Jordan, M. An Introduction to MCMC for Machine Learning. JMLR, vol. 50:5–43, 2003.

[2] Azzopardi, L., Girolami, M., and Risjbergen, K.van. Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In Proceedings of SIGIR, 2003.

[3] Adomavicius, G., and Tuzhilin, A. Towards the next generation of recommender system: A survey of state-of-the-art and possible extensions. IEEE TKDE, 17: 734-749, 2005.

[4] Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet Allocation. JMLR, vol. 3:993-1022, 2003.

[5] Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., and Yu, Y. Optimizing web search using social annotations. In Proceedings of WWW, 2007.

[6] Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. How Do You Feel about "Dancing Queen"? Deriving Mood & Theme Annotations from User Tags. In Proceedings of JCDL, 2009.

[7] Bundschus, M., Yu, S., Tresp, V., Rettinger, A., Dejori, M., and Kriegel, H. Hierarchical Bayesian Models for Collaborative Tagging Systems. In Proceedings of ICDM, 2009

[8] Caverlee, J., Liu, L., and Webb, S. SocialTrust: Tamper-Resilient Trust Establishment in Online Communities. In Proceedings of JCDL, 2008.

[9] Chen, W.Y., Chu, J.C., Luan, J., Bai, H., Wang, Y., and Chang, E.Y. Collaborative filtering for Orkut communities: Discovery of user latent behavior. In Proceedings of WWW, 2009.

[10] Daud, A., Li, J., Zhu, L., and Muhammad, F. Temporal Expert Finding through Generalized Time Topic Modeling. Knowledge Based Systems, Accepted, 2010.

[11] Daud, A., Li, J., Zhu, L., and Muhammad, F. Knowledge Discovery through Directed Probabilistic Topic Models. a Survey. Journal of Frontiers of Computer Science in China (FCS), Accepted, January 2010.

[12] Daud, A., Li, J., Zhu, L., and Muhammad, F. A Generalized Topic Modeling Approach for Maven Search. In APWeb-WAIM , pp. 138–149, 2009.

[13] Golub, K., Jones, C., Matthews, B., Moon, J., Tudhope, D., and Nielsen, M.L. EnTag: Enhancing Social Tagging for Discovery. In Proceedings of JCDL, 2009.

[14] Hofmann, T. Probabilistic Latent Semantic Analysis. In Proceedings of UAI, 1999.

[15] Halpin, H., Robu, V., and Shepherd, H. The complex dynamics of collaborative tagging. In Proceedings of WWW, 2007.

[16] Hunter, J., Khan, I., and Gerber, A. HarvANA – Harvesting Community Tags to Enrich Collection Metadata. In Proceedings of JCDL, 2008.

[17] Jaschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. Tag Recommendations in Folksonomies. In Proceedings of PKDD, 2007.

[18] Li, R., Bao, S., Yu, Y., Fei, B., and Su, Z. Towards effective browsing of large scale social annotations. In Proceedings of WWW, 2007.

[19] Lo, S., and Lin, C. Wmr: a graph-based algorithm for friend recommendation. In Proceedings of WI, 2006.

[20] Mika, P. Ontologies are us: A unified model of social networks and semantics. JoWS, 5(1):5-15, 2007.

[21] Mishne, G. Autotag: a collaborative approach to automated tag assignment for weblog posts. In Proceedings of WWW, 2006.

[22] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth. P. The Author-Topic Model for Authors and Documents. In Proceedings of UAI, 2004.

[23] Sigurbj¨ornsson, B., and Zwol, R.v. Flickr tag recommendation based on collective knowledge. In Proceedings of WWW, 2008.

[24] Sood, S.C., Owsley, S.H., Hammond, K.J., and Birnbaum, L. TagAssist: Automatic tag suggestion for blog posts. In Proceedings of ICWSM, 2007.

[25] Schmitz, C., Hotho, A., J¨aschke, R., and Stumme, G. Mining association rules in Folksonomies. In Proceedings of IFCS, 2006.

[26] Specia, L., and Motta, E. Integrating folksonomies with the semantic web. In Proceedings of ESWC, 2007.

[27] Tang, J., Leung, H.F., Luo, Q., Chen, D., and Gong, J. Towards ontology learning from Folksonomies. In Proceedings of IJCAI, 2009.