User-Characteristics Topic Model

Wenfeng Li, Xiaojie Wang, and Shaowei Jiang

Beijing University of Posts and Telecommunications

Abstract. This paper proposes a method to capture user's characteristics in a topic model frame, where user characteristics act as a latent variable that does not depend on texts. As it is obvious that different people possess different characteristics, they may perform differently even when they are facing the same document. These different characteristics can be showed as different views or different wording preference. We think this phenomenon has a great impact on modeling texts written or labelled by different people, especially on topic modeling. Experiments show that the model with user characteristics outperforms the original models and other similar topic models on corresponding tasks. A combination of the user's characteristics can not only provide better performance on normal topic modeling tasks, but also discover the user's characteristics.

Keywords: user characteristics; topic modeling; personalized model

1 Introduction

With the development of Web2.0, users are becoming more and more deeply involved in the Internet, not only as readers, but also as authors. This development has made the quantity of text corpora on the Internet increase rapidly. As a result, it becomes more and more challenging to organize corpora efficiently, through this, users can find what they need conveniently. Dimensionality reduction is a reasonable way to model large amount of data and get short descriptions for texts which is useful for certain basic tasks such as classification or relevance judgments.

A vast number of statistical learning methods have been used to model the texts. Among them, a series of Latent Dirichlet Allocation (LDA) based topic models initiated by Blei[1] have been developed. LDA uses topics as latent variables for text description. It has been extended in several different ways and has achieved success in some applications. For example, supervised LDA [2] assumes that there is a label generated from each document's topic distribution. Labeled-LDA [3], TagLDA[4] and Multi-Multinomial LDA (MM-LDA)[5] have been used to model multi-labels text. Labeled-LDA constrains the topic distribution by user's labels as supervised information, while the tag set and the word set are assumed to be independently sampled from the document in TagLDA/MM-LDA. The Author-Topic Model (ATM)[6][7] models the interests and topics based relations of authors. The Author Interest Topic Model (AITM)[8] allows a number of possible latent variables to be associated with authors' interests, where the model assumes that each author has only one interest for one document.

Among these models, a basic assumption is all users (including authors and readers) have the same word distributions for a same topic. That is to say, when a topic is found, the word distribution of the topic is same for all users, no matter who they are.

But the fact is that when different people talk about the same topic, there is a big probability they will prefer to use different words. That is to say, for a same topic, there may be several different word distributions for different kind of people. For example, when two people talk about a mobile phone, one may first think about its capability of communication, while the other may first talk about its convenience, depending on their backgrounds and interests. When they talk about the convenience of the mobile phone, one may first use "portable", another may first use "carry-on", because of different wording preferences they may have. It is the same when different people read a document on the same topic, they will be concerned with different words in the document or use different words to tag the documents, depending on their backgrounds and/or wording preferences.

Based on the above observations, we assume word distributions for a topic is not only dependent on the topic, but also dependent on users. We assume there is a latent user characteristics for different groups for people, which makes different groups of users have different word distribution even for a same topic.

This paper aims to capture both latent topics and latent user characteristics in one LDA based model. Due to existing user characteristics, the word distribution on a similar topic for different users will be different. A topic model that does not concern these differences can be thought of as an average model of a large collection of different users with different characteristics. By making use of user-specific differences of topics, we aim to not only achieve better topic modeling for documents, but also extracting more information of both writers and readers(taggers) of the documents .

To combine the user's characteristics in text modeling, we develop a user-characteristics LDA (UC-LDA). The difference between our model and previous is that our model assumes that words of a document are not only rested with document's topic distribution, which is the same as that in LDA, but also controlled by the user's characteristics distribution. Experimental results shown that our model (UC-LDA) outperforms LDA, ATM and AITM significantly in text modeling task. In addition, it can discover some interesting results about user's characteristics which cannot be given by previous ones. Also, we applied the idea of user-characteristics (UC) to TagLDA (UC-TagLDA), from the experimental results, we can find UC is not just specific to a certain topic model, it can be applied to a wide range of topic models.

The organization of this paper is as followings: Section 2 describes how our model works with the example of UC-LDA and UC-TagLDA. The experimental results are shown in Section 3. Section 4 draws some conclusions.

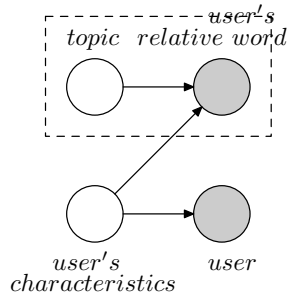
2 User-Characteristics Topic Model

2.1 Motivation

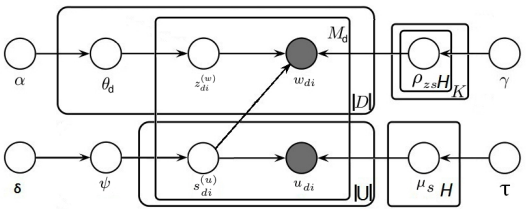
Topic models like LDA only concern the generative process of the documents. For each document in the LDA model, the topic distribution θ_d is a multinomial distribution randomly sampled from a Dirichlet distribution, for each word in document d, the topic assignment $z_{di}^{(w)}$ is chosen from this topic distribution for the ith word, and then a word w_{di} is generated from a topic-specific multinomial distribution $\phi_{z_{di}}$.

As we have argued, when different people talk about the same topic, there is a great probability that they will prefer to use different words. A topic model that does not concern these differences can be thought as an average model of a large group of different users with different characteristics. We now add the user's characteristics to the generative process based on the following assumptions: When a word is chosen for a topic by a user, it not only rests with the topic, but also rests with that user's characteristics. Meanwhile, different types of characteristics generate different users.

These assumptions can be represented in graph shown in Figure 1(a).



(a) Schema of basic assumption behind UC-LDA



(b) Graph representation of UC-LDA

2.2 Description of User-Characteristics LDA(UC-LDA)

UC-LDA can be thought as a combination of the above assumptions and LDA. It gives a way to model users and documents at the same time.

With the combination of Figure 1(a) and LDA, we can get the graph representation of UC-LDA in Figure 1(b). In UC-LDA, each word is not only influenced by its topic assignment, but also by characteristics of the user who is relative to this word. Our notation is summarized in Table 1.

And the generative process of the User-Characteristics Topic Model is shown as follows:

1. Draw H multinomial μ_s from Dirichlet prior τ;
2. Draw $K \times H$ multinomial ρ_{zs} from Dirichlet prior γ to represent the tag distribution, one for each topic z assigned to the characteristics s;
3. Draw a multinomial ψ from Dirichlet prior δ;
4. For each document d, draw a multinomial θ_d from a Dirichlet prior α;
   (a) For each word w_{di} and the user u_{di} who is relative to this word,
      i. Draw a characteristics $s_{di}^{(u)}$ from multinomial ψ, and then draw a user from $\mu_{s_{di}^{(u)}}$
      ii. Draw a topic $z_{di}^{(w)}$ from multinomial θ_d, and then draw a word from $\rho_{z_{di}^{(w)} s_{di}^{(u)}}$

Table 1. Notation used in our model

SYMBOL	DESCRIPTION
D,$\|D\|$,d	D is a collection of documents, $\|D\|$ is the number of documents in the collection, and d is a document in the collection.
W,$\|W\|$,w	W is a collection of word tokens, $\|W\|$ is the number of tokens in the collection, and w is a word in the collection.
T,$\|T\|$,t	T is a collection of tag tokens, $\|T\|$ is the number of tokens in the collection, t is a token in the collection.
U,$\|U\|$,u	U is a collection of users, $\|U\|$ is the number of users, and u is a user in the collection
w_{di}	the ith word in document d
t_{dj}	the jth tag in document d
u_{dj}	the user who give the jth tag in document d
K	number of topics
H	number of user's characteristics
N_d	number of words in document d
M_d	number of tags in document d
$z_{di}^{(w)}$	the topic assigned to the ith word in the document d
$z_{dj}^{(t)}$	the topic assigned to the jth tag in the document d
$s_{dj}^{(u)}$	the characteristic assigned to user who give the jth tag in the document d
θ_d	the topic distribution of document d
ψ	the characteristics distribution on the corpus
ϕ_z	the word distribution of topic z
ρ_{zs}	the tag distribution of topic z specific to characteristic s
μ_s	the user distribution of characteristic s

2.3 Inference

As a topic model can not be exactly inferred, we use Gibbs sampling to get an approximate inference of our model. For each iteration, we need to sample the

topic of each word, and also need to sample the characteristics of the user who gives that tag.

The Gibbs sampling procedure can be seen in Figure 1. Where,

```
for each iteration:
  for d in D:
    for i = 1 to N_d:
      draw z_di^(w) from p(z_di^(w)|·)
      draw s_di^(u) from p(s_di^(u)|·)
      update n(z_di^(w), s_di^(u), w_di), n(s_di^(u), u_di) and n^(w)(d, z_di^(w))
    end for
  end for
end for
```

Fig. 1. Gibbs sampling process of UC-LDA

$$p(z_{di}^{(w)} = z|\cdot) \propto \frac{n(z, s_{di}^{(u)}, w_{di})_{\neg t_{di}} + \gamma}{\sum_t (n(z, s_{di}^{(u)}, t) + \gamma) - 1} \times \frac{n^{(w)}(d, z)_{\neg w_{di}} + n^{(w)}(d, z) + \alpha}{\sum_d (N_d + M_d) + \alpha K - 1} \quad (1)$$

$$\begin{aligned} p(s_{di}^{(u)} = s|\cdot) \propto & \frac{n(s, u_{di})_{\neg u_{di}} + \tau}{\sum_u (n(s, u) + \tau) - 1} \times \frac{\sum_u n(s, u_{di})_{\neg u_{di}} + \delta}{\sum_d M_d + \delta H - 1} \\ & \times \frac{n^{(w)}(d, z_{di}^{(w)})_{\neg t_{di}} + n^{(w)}(d, z_{di}^{(w)}) + \alpha}{\sum_d (N_d + M_d) + \alpha K - 1} \end{aligned} \quad (2)$$

And, $n(z, s, w)$ is the number of tokens of word w is assigned to topic z with user assigned to characteristic s, $n^{(w)}(d, z)$ is the number of word tokens in document d is assigned to topic z, and $n(s, u)$ is the number of occurrence of user u is assigned to characteristic s.

2.4 UC-TagLDA

User characteristics can also be combined with other topic models. We describe the combination of User characteristics with TagLDA in this section.

TagLDA is used to model social tagged data like del.icio.us. In such a social tagged system, different user may use different tags to tag a same content since they will concern different aspects of the content or they may use different words to describe the same content. In this way, we add the users characteristics to the generative process based on the following assumptions: when a tag is chosen for a topic by a user, it not only rests with that topic, but also rests with the user's characteristics. Meanwhile, different types of characteristics generate different users.

To model the tagged social documents, TagLDA considers the generative process of both words and tags. For document d, the ith word w_{di} is generated in the same way as in LDA, while topic assignment $z_{dj}^{(t)}$ for the jth tag is chosen from the document's topic distribution θ_d, and that tag t_{dj} is also generated from a topic-specific multinomial distribution $\rho_{z_{dj}}$. In UC-TagLDA, the difference is tag t_{dj} is not only influenced by tag distribution $\rho_{z_{dj}}$, but also influenced by that user's characteristics assignment $s_{dj}^{(u)}$ which also influence the generation of user u_{dj}.

The graphical model is shown in Figure 2. And we also use Gibbs sampling for posterior inference.

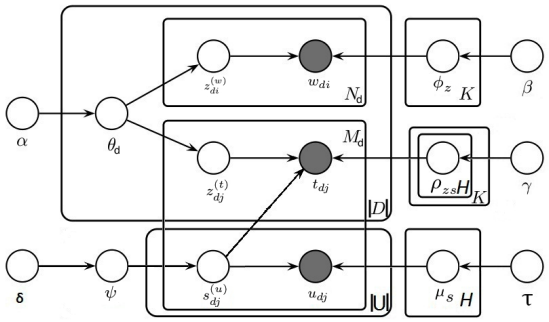


Fig. 2. Graph representation of UC-TagLDA

3 Experiments

3.1 Dataset

UC-LDA We use two dataset of papers for experiments about UC-LDA, the abstract from CiteseerX and the full paper from NIPS. The CiteseerX dataset (1.2GB xml file) is downloaded from CiteSeerX OAI collection and it has 456,353 abstracts from 650,478 authors. The NIPS dataset (35MB mat file) [9] has 2,484 papers from 2,865 authors.

We randomly choose about 90% as the training data and about 10% as the testing data. For CiteseerX dataset, there are 407,812 training documents (598,745 authors) and 48,541 testing documents (60,1973 new authors). For NIPS dataset, there are 2,207 documents (2,290 authors) and 277 documents (575 new authors) left for testing.

UC-TagLDA We use the data from del.icio.us provided by DAI Labor[10] for UC-TagLDA. Del.icio.us is a social bookmarking system in which users can tag

each of their bookmarks freely. Each record of the data consists of three parts, including user, url and tags.

We chose the bookmarks collected in 2004 for the experiments. To avoid data sparseness, we have removed the URLs that have been tagged by fewer than 20 users, and also removed those users who tagged fewer than 50 URLs. After preprocessing, 1121 users and 2476 URLs remained. We then crawled the web pages of these URLs. After preprocessing to remove the irrelevant content and the web page stop words , there were about 143 words left for each page.

As the dataset used for UC-TagLDA is small, we ran the experiments using this dataset with 10-crossfold validations.

3.2 Perplexity

UC-LDA In the comparison experiments for UC-LDA, we compute the perplexity of words comparing among LDA, ATM, AITM and UC-LDA on 10, 20, 50, 100, 150, 200, 500 and 1000 topics. where 1000 topics only set for NIPS dataset. As CiteseerX dataset is much larger than NIPS dataset and the limitation of memory, it is hard to do experiments with exactly the same number of topics on both dataset. For AITM, we also set different interest number (1, 5, 10, 20, 50, 100, 150, 200), and choose the best result (lowest perplexity) on each topic number for comparative purposes.

The experimental results of NIPS and CiteseerX are respectively shown in Table 2 and Table 3.

Table 2. UC-LDA:Perplexity on NIPS Dataset

Topic Num	LDA	ATM	AITM	UC-LDA				
				1 cha.	5 cha.	10 cha.	20 cha.	50 cha.
10	1985.94	4079.68	3242.43	1992.67	1774.21	1576.88	**1428.29**	1639.83
20	1770.34	4271.98	3242.24	1765.18	1714.58	1597.98	**1415.06**	1864.60
50	1516.10	4509.35	3242.34	1532.38	1494.18	**1339.65**	1879.24	2351.45
100	1370.35	4526.52	2991.91	1381.87	1334.28	**1284.41**	1573.89	2011.06
150	1304.43	4588.58	3107.88	1318.61	**1205.48**	1666.07	2147.20	2613.61
200	1349.78	4540.07	3242.56	**1277.93**	1636.87	1926.90	2256.08	2690.94
500	**1337.54**	4566.81	3242.55	1338.06	1955.98	2319.86	2962.63	3163.50
1000	1447.76	4597.52	3242.94	**1438.19**	2482.52	2942.73	3599.60	4577.21

Table 2 shows our model outperforms the other models on 10, 20, 50,100 and 150 topics. And our model (with 1 characteristics) and LDA has nearly the same perplexity on 200, 500, 1000 topics,

When the topic number is set to a small value, user characteristics can separate a topic into several sub-topics by considering user difference, and cause better performance. As the increasing of number of topic, LDA gets its best

performance, but the seperation of topics in LDA and UC-LDA are not same. The best perplexity of UC-LDA is got at 5 characteristics and 150 topics, while LDA does not get best perplexity at $5 \times 150 = 750$ topics. The best performance of UC-LDA improves about 9.9% comparing with best LDA.

And this results can also show that LDA is a special case of UC-LDA with 1 characteristic.

From Table 3, we can get the same conclusion as in Table 2. The best performance of UC-LDA improves about 12.2% comparing with best LDA. UC-LDA brings bigger improvement of perplexity on bigger data.

Table 3. UC-LDA:Perplexity on CiteseerX Dataset

Topic Num	LDA	ATM	AITM	UC-LDA				
				1 cha.	5 cha.	10 cha.	20 cha.	50 cha.
10	1638.68	2453.72	2321.70	1659.38	1495.68	1320.19	1196.02	**1118.37**
20	1572.30	2758.42	1984.76	1603.35	1528.15	1457.65	**1200.29**	1211.38
50	1328.74	2786.09	2089.28	1322.25	1329.15	**1248.31**	1370.23	1545.94
100	1280.06	2664.23	1972.95	1281.47	**1204.51**	1452.76	1621.43	1892.18
150	1274.46	2772.25	2106.21	1270.66	**1264.15**	1729.40	1925.38	2229.91
200	1306.67	2836.97	2101.23	**1287.89**	1485.95	1875.92	2381.29	3095.13
750	1315.41	-	-	-	-	-	-	-

UC-TagLDA In UC-TagLDA, we have a latent variable for the user's characteristics in addition to the latent variable for topics common in both of our model and TagLDA. For topics, we tested six different values. For user's characteristics, we tested three different values. The experimental results are shown in Table 4 and Table 5.

Table 4 shows our model has smaller perplexities than those in TagLDA on 5 and 10 characteristics with different topics number. The best perplexity of UC-TagLDA improves about 1.4% comparing with best TagLDA. Perplexities on tags shown in Table 5 bring us to the same conclusion. Our model outperforms TagLDA and the best perplexity of our model improves about 18.8% comparing with best TagLDA.

Unlike TagLDA, ATM models authors and documents at same time. We therefore compare our model to ATM.

To fit the data requirements of ATM, we mix the tags and users of each URL together and assume that all URL tags were co-created by all users who tagged the URL. The same parameters are set as the previous experiments.

Experimental results are shown in Table 5. It shows that UC-TagLDA significantly outperforms the ATM model on social tagged data.

Table 4. UC-TagLDA:Perplexity of words on del.icio.us

Topic Num	Tag-LDA	UC-TagLDA		
		5 cha.	10 cha.	20 cha.
10	3896.92	**3805.34**	3930.83	3898.23
20	3285.39	**3019.31**	3293.35	3283.66
30	2895.01	**2812.14**	2838.54	2852.27
50	2874.11	2811.62	**2801.13**	2847.21
100	2775.25	2742.02	2740.11	**2736.03**
200	2798.67	2751.30	**2739.42**	2784.93

Table 5. UC-TagLDA:Perplexity of tags on del.icio.us

Topic Num	Tag-LDA	UC-TagLDA		ATM
		5 cha.	10 cha.	
10	165.23	**144.87**	158.24	1008.43
20	113.24	**97.95**	116.26	613.34
30	95.6	**87.61**	92.88	895.54
50	117.4	**83.25**	94.39	1032.57
100	92.96	**70.62**	86.52	1175.91
200	87.02	**71.23**	85.29	1264.21

3.3 Word distributions over different characteristics

The user's characteristics can perform in different forms. Some characteristics may represent different aspects of a topic, and some characteristics may represent different wording preferences. Table 6 and Table 7 respectively show the word/tag rank of UC-LDA(on Citeseerx dataset) and UC-TagLDA.

In Table 6, for each topic, we list the word rank (Top-10) of two different characteristics. And we can find that although the top-10 words are nearly in the same set, they are in different order, which demonstrats different wording preferences. For example, in topic-7 (CiteSeerX), top 9 words are the same but in different position in charac-0 and charac-4. "medical" is the top one word in charac-0, and the second word in charac-4, "patients" is the second word in charac-0, and the 7th word in charac-4. "neural" is the second word in charac-0, and the 5th word in charac-4. Obviously, these different topics cause by user wording preference can not found by LDA models. And we can also find the different perspective of the same topic. For example, topic-4 (NIPS) is about physiological, where charac-9 is interested in the physiological property, while charac-15 concerns more technical details.

In Table 7, topic-29 has to do with the web service, but charac-2 users are concerned with the usage of blog, while the charac-5 users may be paying more attention to the search technology. For topic-47, charac-0 and charac-3 are both interested in the hardware product, because they have used almost the same

most used words, but in different order when they talk about a same topic, demonstrating different wording preferences.

Table 6. UC-LDA: The Top-10 words of each characteristics for the same topic (50 topics, 20 characs)

Topic-7(CiteseerX)		Topic-9(CiteseerX)	
charac-0	charac-4	charac-0	charac-6
medical	brain	image	image
patients	medical	images	images
neural	clinical	visual	objects
clinical	activity	objects	video
brain	neural	motion	visual
patient	diagnosis	video	object
diagnosis	patients	objects	motion
activity	patient	spatial	spatial
treatment	treatment	robot	shape
cortex	disease	tracking	robot
Topic-4 (NIPS)		Topic-19 (NIPS)	
charac-9	charac-15	charac-6	charac-19
factorizations	volatile	table	contents
earliness	workshop	contents	table
rheological	detect	list	tables
forthcoming	division	figure	ftp
nanomaterials	renovation	tables	list
mell	eaor	preface	introduction
lipoproteins	electromyogram	introduction	figures
offending	mqp	postscript	esi
locationaware	neurotransmitter	ftp	acknowledgements
incidences	closures	acknowledgements	preface

3.4 Application on recommendation(UC-TagLDA only)

Both UC-TagLDA and TagLDA can be used for tag recommendation. As [5] shows TagLDA has better performance compared to K-means on tag recommendation, we designed two groups of experiments to compare their recommendation performance.

In the experiments, the topic number is set to 50 for both TagLDA and UC-TagLDA, for UC-TagLDA, the number of user's characteristics is also set to two different values, 5 and 30.

For each webpage, we chose top-N tags as the recommended tags and compared them to the user's tags.

To evaluate the model, we randomly selected 90% of 2476 URLs as training data, and the remaining was used as test data.

Table 7. UC-TagLDA: The Top-10 tags of each characteristics for the same topic on del.icio.us(50 topic, 30 characs)

Topic-29		Topic-47	
charac-2	charac-5	charac-0	charac-3
wisdom	copyright	hardware	shopping
semantic	seo	shopping	hardware
history	searchengine	tech	gadgets
tricks	hardware	diy	tech
wordpress	ajax	geek	hacks
article	info	cool	diy
random	amusements	howto	geek
webdizajn	im	gadgets	gadget
java	commerce	video	shop
proxy	html	technology	technology

Table 8 gives the results of precision, recall and F1 score on our preprocessed data from DAI-Labor dataset.

Since we have more parameters in UC-TagLDA than in TagLDA, we wonder if the scale of training data will bring different effects. To evaluate the influences on the amount of training data, we randomly chose 1/2, 1/4 and 1/8 of the whole training data to train the model, and evaluated on the same set of testing data. The experimental data is shown in Table 9. In the experiment, the topic number was also set to be 50, and the evaluation is on the top 10 tags.

Based on the above experiments, we ranked the tags of each user on the web page for each model, and calculated the average ranks of the user's real tags in the models. The ranking results are shown in Table 10, and from the results, it is easily to see that our model significantly outperforms TagLDA.

4 Conclusions

This paper proposes a new idea of topic model to address the problem of user characteristics. User's relevant words are assumed to be not only generated from latent topics as in a normal topic model, but also influenced by the user's characteristics. Experimental results show that the model with user's characteristics(UC-LDA & UC-TagLDA) outperforms the previous models(LDA & TagLDA) and other similar topic models(ATM & AITM) on text modeling.

Furthermore, our model can also give some interesting results about user characteristics. We have found two varieties of user characteristics from our model, one concerns different views of a topic, and the other is different wording preference.

Table 8. The evaluation result of recommendation

Topic Num	Method	Recall(%)	Precision(%)	F1 score
2	UC-TagLDA(5 cha.)	**10.44**	**11.01**	**0.1072**
	UC-TagLDA(30 cha.)	10.35	10.92	0.1063
	TagLDA	9.32	9.83	0.0957
5	UC-TagLDA(5 cha.)	16.20	6.83	0.0961
	UC-TagLDA(30 cha.)	**18.12**	**7.64**	**0.1075**
	TagLDA	15.65	6.59	0.0928
10	UC-TagLDA(5 cha.)	23.52	4.96	0.0819
	UC-TagLDA(30 cha.)	**24.86**	**5.24**	**0.0866**
	TagLDA	21.42	4.52	0.0746
20	UC-TagLDA(5 cha.)	31.92	3.36	0.0608
	UC-TagLDA(30 cha.)	**32.77**	**3.55**	**0.0641**
	TagLDA	30.35	3.19	0.0578

Table 9. The evaluation result of recommendation with different amounts of training data

Prop.	Method	Recall(%)	Precision(%)	F1 score
1/8	UC-TagLDA(5 cha.)	24.17	5.09	0.0841
	UC-TagLDA(30 cha.)	**24.98**	**5.26**	**0.0870**
	TagLDA	18.32	4.41	0.0711
1/4	UC-TagLDA(5 cha.)	24.42	5.22	0.0862
	UC-TagLDA(30 cha.)	**24.93**	**5.25**	**0.0868**
	TagLDA	21.09	4.45	0.0735
1/2	UC-TagLDA(5 cha.)	**23.99**	**5.06**	**0.0835**
	UC-TagLDA(30 cha.)	23.88	5.04	0.0832
	TagLDA	21.06	4.44	0.0734
1	UC-TagLDA(5 cha.)	23.52	4.96	0.0819
	UC-TagLDA(30 cha.)	**24.86**	**5.24**	**0.0866**
	TagLDA	21.43	4.52	0.0746

Table 10. Ranking of user's tags

Model	UC-TagLDA		TagLDA
	5 cha.	30 cha.	
Ave. rank	181.37	157.25	212.91

5 Acknowledgements

The work presented in this paper is supported by the National Science Foundation of china (No. 61273365) and National High Technology Research and Development Program of China (No. 2012AA011104).

References

1. D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.
2. D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *NIPS*, 2007.
3. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.
4. X. Si and M. Sun, "Tag-lda for scalable and realtime tag recommendation," *Journal of Computational Information Systems*, 2009.
5. D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina, "Clustering the tagged web," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ser. WSDM '09. New York, NY, USA: ACM, 2009, pp. 54–63.
6. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," *In 20th Conference on Uncertainty in Artificial Intelligence*, 2004.
7. M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Groffiths, "Probabilistic author-topic models for information discovery," in *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 306–315.
8. N. Kawamae, "Author interest topic model," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 887–888.
9. A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," *The Journal of Machine Learning Research*, vol. 8, pp. 2265–2295, 2007.
10. R. Wetzker, C. Zimmermann, and C. Bauckhage, "Analyzing social bookmarking systems:a del.icio.us cookbook," *European Conference on Artificial Intelligence (ECAI)*, 2008.
11. T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.
12. D. M. Blei and J. D. Lafferty, "Correlated topic models," *NIPS*, p. 147, 2006.