An Improved Lda Model in Micro-Blog Tags Extracting Based on Multi-Tags

Jianfang Wang[1], Kunxiao Shen[2], Anfeng Xu[1] and Yihua Lan[1,*]

[1]*School of Computer and Information Technology, Nanyang Normal University, Nanyang, 473000, China;* [2]*Biological Engineering Master Graduate Student, Nanyang Normal University, Nanyang, 473000, China*

Abstract: This article mainly discusses how to extract the interested information from massive amounts of micro-blogs and recommend right information to user, which is a hot research area in recommendation systems and social networks, too. To solve this problem, a model called Multi-tags Latent Dirichlet Allocation is proposed. Using this model, topics paid attention by users can be mined effectively and the defect of low degree of differentiation for the short blog content is settled. Experiments showed that the tags of user's micro-blog can be figured out with this model which makes users manage their resources at their convenience and others find their needed resources through tags. The results, experimented on real micro-blog data set, indicate that this model works better than traditional model on extracting tags. Standard measuring index Perplexity is applied to this model to estimate the likelihood of new text. If the number of topics is selected appropriately, the accuracy will be raised to almost 10%.

Keywords: LDA model, micro-blog tags, MTLDA, resource sharing.

1. INTRODUCTION

Recent years, with the rise of social networks such as Facebook, Twitter, Weibo, etc, various kinds of information is generated, broadcasted and consumed constantly by all kinds of crowd. Social network has become the main distributing centre of information and is filled with huge number of information, personality activities and large social networks. How to extract the interested information from massive amounts of micro-blogs and recommend right information to user is hot research area in recommendation systems and social networks. The personalized information expressed by user tags makes others find him more exactly and enriches the recommendation information to friends. So how to model user's interest and recommend suitable Weibo tags is the main research content of this paper.

2. MATERIALS

The tags of micro-blog user are the characteristics that others can find him more precisely and enrich the recommendation information to friends. But most users have no tags for their micro-blogs. They don't add tags when writing blogs, or add little tags, or the added tags are not suitable for the blog content, which impede information sharing to some extent [1, 2].

LSA (Latent Semantic Analysis) is a novel method proposed by Deerwester S, *et al.* [3] on mining text topic by linear algebra. Dimensionality reduction method of SVD (Singular Value Decomposition) is used here to mine the latent semantic structure in document. Query and analysis on relativity are carried out in lower dimensional space. Through singular value decomposing, the latent relativity can be dig out better [4]. Researches show that when the dimension of this semantic space is similar to that of human semantic understanding, LSA is more similar to human's understanding and surface information is transformed into deep-seated abstraction [4, 5].

PLSA (Probabilistic Latent Semantic Analysis) is a probability model proposed by Kim Y M [6] whose research is based on LSA. This model is based on Maximum Likelihood and generative models. Following the dimensional reduction idea of LSA, PLSA adopt that text is a high-dimensional data in common text expression (TF-IDF) and that a limited number text topic mining in high-dimensional space is mapped into lower-dimensional semantic space through dimension reduction. Dirichlet prior distribution is added into LDA (Latent Dirichlet Allocation), which is a breakthrough extension of PLSA. Blei [4], introducer of LDA, indicated that uniform probability model was not used in calculating topic probability corresponding to document. Excessive amounts of parameters would result in over fitting and it was hard to allocate probability to the documents outside the training set. To overcome this drawback, hyper-parameter is introduced to LDA to form a three layers Bayesian model of document-topic-words [7]. Model is deduced through probability to find semantic structure of document set and dig out the topic of the text. At present, LDA model has become the criteria of topic modeling and applied in many areas, especially in the research scope of social networks and social media [8, 9]. It has excellent prospects for research and application and has great potential in Weibo topic mining. If being improved, it can be applied to social networks better.

This paper proposes a new improved method for extracting micro-blog tags, named MTLDA (Multi-Tags LDA), based on the LDA topic model through analyzing the characteristics of Weibo user.

3. FUNDAMENTALS OF ALGORITHMIC

Social tags system usually includes three primary elements--user, resource and tags, thus folksonomies set is formed as following:

$$F = \{User, \mathrm{Re}source, Tag_1, Tag_2, \cdots, Tag_n\} \qquad (1)$$

In this problem, there is no resource only user and tags set in Weibo data [10]. If we only use these two sets to recommend micro-blog topics, the result set will deviate from the user's real interest due to the short contents of blogs. So the historical blog data combined with current blogs can be regarded as the user's content, and the topic distribution can be easy to compute. At the same time, because users often pay attentions to others having same interests, friends are more similar to the user than followers in hobbies and interests. Friends set is added into the analysis model to restrain the topic model which modifies the user-resource-tags set to user-friends-tags set, a constraint LDA topic model. The relationships matrix of these three sets is set up: user-friends matrix M, friends-tags matrix M', user-tags matrix m''. In matrix M, if user U_i has paid attention to some friend, the corresponding m_{ij}=1, otherwise, m_{ij}=0. In matrix M', if a friend A_i is marked with tag t_k, then m't_k=1; if not, m't_k=0. In matrix M'', whether user U_i has chosen tag t_k needs to be judged, if done, m_{tk}''=1; otherwise mtk''=0.The similarity or distance between all relationship groups (User, Attentions, Tag_1, Tag_2,⋯Tag_n) is calculated to recommend tags.

4. ORIGINAL LDA TOPIC MODEL

Each document in corpus generated by LDA model can contain several topics. The generative process is shown as follows:

Choose parameter θ ~ p(θ);

For each of the N words wn;

Choose a topic zn ~ p(z|θ);

Choose a word wn ~ p(w|z);

Here, θ is a topic vector whose column represents the probability of every topic accruing in document and it is a nonnegative normalized vector. P(θ) is the Dirichlet distribution of θ. N denotes the number of document's words. w_n is the Nth generated word named w. p(w) is the distribution of w that can be obtained by statistical learning in corps. For example, if a book is given, statistics will be done on probability of every word in it. P(z|θ) expresses the probability distribution of topic z when θ is given and its value is θ, that is, p(z=i|θ)= θ. p(w|z) is the distribution of w when z is given and it can be seen as a matrix of k*V, in which k is the number of topics, and V is the number of words, and every line of the matrix means probability distribution of words corresponding to the topics or the probability of each word contained in topics z. Each word is generated on the basis of a certain probability. Diagram of this model is shown in Fig. (**1**).

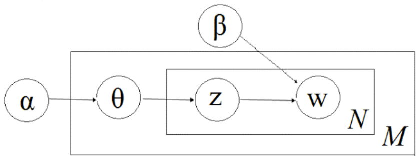

Fig. (1). Original LDA topic model.

A topic vector named θ is first selected in this model to determine the probability of each selected topic. Here θ is a variable in the document level and there must be a θ corresponds to a document, that is to say, θ is sampled to each generated document for the reason that the probabilities of topic z is different. Then a topic z is selected in topic distribution vector θ to generate a word w on basis of the probability distribution of topic z. Here, z and w are both variables of word level and z is produced by θ, w is produced by z and β. A word w corresponds to a topic z.

The united probability of LDA is shown here:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \qquad (2)$$

The main task of LDA model is to learn training two control parameters, α and β, from given input corpus to determine model for generating document. Here α is a vector parameter of p(θ), or parameter of Dirichlet distribution, to generate a topic vector θ and β is a word probability distribution matrix p(w|z) corresponding to each topic. For α and β are parameters at the corpus level, namely, it is the same for each document, sampling is done only once during the generation process.

In the original LDA topic model shown above, every document serves as a probability distribution of some topics and every topic as a probability distribution of many words. Assume that number to topic z_n of every word is calculated for any document d in document set consisting of D document and each topic is extracted from topic set composed of k topics, probability distribution of each document on the topic and the words in each topic can be eventually computed by iteration.

5. IMPROVED MTLDA TOPIC MODEL

Because micro-blog is often limited to 140 words, the probability of the same word occurred in two short documents is too small to measure their similarity. At the same time, for most users one of the micro-blog's functions is to get information and friends represent the topics that users interested in. The relationship between followers and user is more distant than that between friends and user, for paying attention to someone is an initiative choice and being paid attention by someone is a passive choice. Therefore, in our analysis model, friends set is added to restrain the topic model and the model of user-resource-tags set is modified to user-friends-tags set which is named constraint multi-tags LDA(MTLDA) topic model. The entire historical document of user is viewed as the present user's document for the lack of original LDA. The micro-blog's content is divided based on user and all tweets posted by user is merged to stand for user. Then the standard three layers LDA model of document-topic-word is changed into user-topic-word model. So friends' information of user can be acted as constraint when extracting tags from tweets. The improved LDA model is shown in Fig. (**2**).

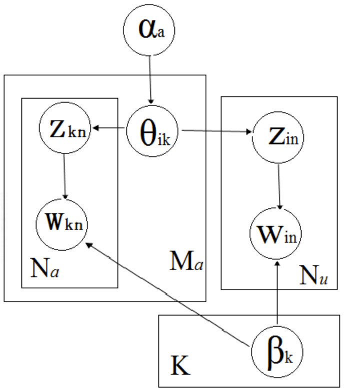

Fig. (2). Improved constraint MTLDA model.

In this LDA model, N*u* is the total number of words in the user's Weibo set similarly, N*a* is friend's Weibo set. M*a* is a set of friends in which each friend and user together forms a friend pair of (u,v). At first, topic distribution vector θ_{ik} is produced with the vector α_a. Then for the number of N*a* tags of friends, a word w_{kn} is produced according to the word probability distribution of topic z_{kn} selected from θ_{ik} combined with corresponding β_k matrix. It is the same to each user in M*a* set. Last, for the number of N*u* tags of user, a user's word w_{in} is also produced according to the word probability distribution of topic z_{in} combined with β_k. Here z_{in} is selected from friend's Weibo θ_{ik}.

6. MTLDA ALGORITHM DESCRIPTION

The calculating process above can be described as following:

For each pair of (u,v) constituted of friend and user:

1) According to α, produce $\theta_{ik} \in \Delta(K)\text{-Dir}(.|\alpha_a)$

2) To attention friend's Na tags, generate $z_{kn} \in \{1,2,\ldots\ldots,K\}\text{-Mult}(.|\theta_{ik})$, and build $w_{kn} \in \{1,2, \ldots\ldots,V\}\text{-Mult}(.|\beta_k)$.

3) To user's Ma tags, make $z_{in} \in \{1,2,\ldots\ldots,K\}\text{-Mult}(.|\theta_{ik})$,and $w_{in} \in \{1,2, \ldots\ldots,V\}\text{-Mult}(.|\beta_k)$.

The tags used in this algorithm are the annotation of user's demanded and interested resources. They are added by users themselves to the resources. These description tags represent hobbits of user which expand the relationship in the traditional recommended system from binary relationship among user and item to ternary relationship among user, item and resource. Tags make users to manage their resources at their convenience and others to find their needed recourses which mean resources sharing.

7. EXPERIMENTAL PROCESS

According to above algorithm, experiment is divided into two steps. In the first step, user's information and Weibo data are collected with Sina Weibo API and pre-processed. In the second step, tags are getting from experiment which is performed on the MTLDA model proposed in this paper compared with traditional LDA topic model.

Authenticated user's information is obtained from 8 distinct popular areas based on recommended user interface. They are science and technology, anime, entertainment, healthy, physical training, cars, property and mass media. In addition, latest Weibo of user are get no more than 300 pieces taking advantage of user's Weibo list interface. All of these data coming from internet should be pre-processed for their noise. Some pieces are removed with less than 10 number of replies and forwards. 50 users are selected from every area based on the real effective data. Similarly, users whose Weibo is less than 10 are filtered out and 400 users are passed into the experimental data set. The final data used in the experiment is described in Table **1**.

8. EXPERIMENTAL RESULTS

To compare the accuracy of different methods, tags are extracted with LDA model in our experiment which is modified with GibbsLDA++ toolkit to adapt the extracting mission and parameters are experimental values referenced to Weng J, *et al.* [11], in which α=50/T (T is the number of topics) and β=0.01. The number of topics is set to 8-15. At first, tags are extracted with LDA model. The historical Weibo content posted by user represents user himself. So that, document-topics-words is set up as three layers of LDA model and topics distribution vector θ_{ik} is derived from this model combined with α_a vector parameter, as is shown in column 2 of Table **2**.

Then for the friends' tags, a topic z_{kn} is selected from θ_{ik}. $z_{kn} \in \{1,2,\ldots\ldots,K\}\text{-Mult}(.|\theta_{ik})$ is produced according to the probability distribution of topic z_{kn}. Then a word $w_{kn} \in \{1,2, \ldots\ldots,V\}\text{-Mult}(.|\beta_k)$ is created combined with corresponding word probability distribution matrix β_k. In the same way, tag words are solved for all of the other friends. At last the user's tag words is obtained by viewing all of the tags sets as constraint to the user's tags, which is also shown in Table **2** as column 3. Experiments show that the tags obtained by MTLDA model proposed in this paper is more various than that by LDA and is more suitable for user's interesting. This model possesses high distinguish degree.

9. DISCUSSIONS

To evaluate the performance of MTLDA model and its algorithm, the accuracy of tag recommendation is measured in this paper. Perplexity is a standard measuring index for assessing the performance of language generation model. The result value represents the likelihood estimation of new text in testing set generated with model. The smaller the value, the higher the estimation, in other words, the generating performance of this model is better. The computation formula is shown as the following:

In the formula, U_{test} is the users in testing set, N is the total number of users, w_{ui} is the word set contained in the user's blog w_{ui}, $p(w_{ui})$ is the generating probability of u_i under the guidance of user model, N_{ui} is the total number of words in u_i blog set [12, 13]. The results of this experiment are shown in Fig. (**3**).

Tabel 1. Distribution of experimental data.

Area	Number of Users	Number of Effective Weibo
Science and technology	50	22432
Anime	50	15388
Entertainment	50	24890
Healthy	50	20411
Physical training	50	20866
Cars	50	23945
Property	50	20130
Mass media	50	20743
Total	400	168805

Table 2. Tag words of Weibo extracted by two models

User ID	LDA Model	MTLDA Model
35****7684	Travel, delicacy, music, after 80s	Journey, foodie, delicious, music, movie, listening to songs
56****9031	Movie, Music Awards, wifi, football	The sound of China, happiness, individual, beer, carve out
20****3454	Data mining, technology, internet, Information processing	Data mining, internet, big data, social marketing, artificial intelligence, signal processing, science and technology
15****7230	Music, classical, purple, environment art design	Reading, listening songs, music, nostalgia, inttoversion, sign

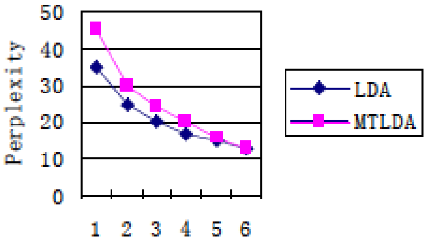

Fig. (3). Accuracy of LDA and MTLDA model.

The results, experimented on real Weibo data set, indicate that MTLDA model has better affectivity than traditional LDA model on extracting tags. If the number of topics is selected appropriately, the accuracy will be raised to almost 10%.

CONCLUSION

The tags of micro-blog provide important reference for friends' recommendation, information recommendation and advertisement delivery. Aiming at the lacking of marked tags, MTLDA model is proposed in this article based on the multi-tags LDA topics model in which the historical blogs set represent the user himself, mass data is classified by LDA model combined with friends, the repeated keywords in the blogs is extracted as the multi-tags of user. Topics paid attention by users can be mined with MTLDA model which solves the problem of defect of low degree of differentiation for the short blog content. Experiments show that the tags of user's micro-blog can be figured out with this model effectively which makes users manage their resources at their convenience and makes others find their needed resources through tags. Resources will be shared among users. In the future, this model can be applied into personality information recommendation to enhance the accuracy and efficiency of information recommendation.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This study was supported in part by the National Natural Science Foundation of China (Grant No. 61401242), the Research Foundation for Advanced Talents of Nanyang Normal University (ZX2014058), Technology Research and Development Program of Lianyungang (Grant No.SH1223), the Research plan for Basic and frontier Research of Henan (Grant No. 142300410044), and the Key Programs of Education Department of Henan Province (Grant No. 14A520057).

REFERENCES

[1] A.Plangprasopchok, J. H. Kang, and K. Lerman, "Analyzing micro-blogs with affinity propagation," *Proc. of the 1st Workshop on Social Media Analytics*, pp. 67-70, 2010.

[2] D. Wu, and R. Xu, "Survey of Clustering Algorithms", *IEEE Trans. on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[3] S.Deerwester, S. Dumais, T. Landauer, "Latent semantic analysis for multiple-type interrelated data objects," *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 236-243, 2006.

[4] D.Blei, "Probabilistic topic models", *Communications of the ACM*, 55(4), pp. 77-84, 2012.

[5] H. Hirsh, and S. Zelikovitz, "Using LSI for Text Classification in the Presence of Background Text," *Proc. of the 10th International Conference on Information and Knowledge Management*, pp. 113-118, 2001.

[6] Y. M. Kim, "An Extension of PLSA for Document Clustering," *Proc. of the 17th ACM Conference on Information and Knowledge Management*, pp. 1345-1346, 2008.

[7] C. C. Yang, and X. N. Tang, "TUT: A statistical model for detecting trends, topics and user interests in social media," *Proc. of the 21st ACM International Conference on Information and Knowledge Management*, pp. 972-981, 2012.

[8] X. Wei, and W. B. Croft, "LDA-based document models for ad hoc retrieval," *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178-185, 2006.

[9] E. P. Lim, J. Weng, and J. Jiang, "Twitter Rank: finding topic-sensitive influential twitters," P*roceeding of the third ACM International Conference on Web Search and Data Mining*, 13, 2013.

[10] B. Zhang, M. Ostendorf, and W. Wu, "Automatic generation of personalized annotation tags for Twitter users," *Human Language Technologies: The 2013 Annual Conference of the Nortti American Chapter of the Association for Computational Linguistics*, HLT'13, 2013.

[11] J. Weng, P. Lime, and Q. He, "What do people want in microblogs? Measuring interestingness of hashtags in twitter," *Proceedings of the 2012 IEEE International Conference on Data Mining*, ICDM'12, 2012.

[12] A. Y. Ng, D. M. Blei, and M. I. Jordan, "Latent Dirichlet allocation", *The Journal of Mach. Learning Research*, pp. 993-1022, 2003.

[13] A. Rappoport, D. Davidiv, O. Tsur, "Enhanced sentiment learning using twitter hash-tags and smileys", *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China. pp, 241-249, 2010.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014