Journal of Theoretical and Applied Information Technology
10th February 2013. Vol. 48 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MODELING USER INTERESTS USING TOPIC MODEL

[1] QINJIAO MAO, BOQIN FENG, [2] SHANLIANG PAN

[1] The School of Electronic and Information Engineering, Xi'an Jiaotong University, 710049,China

[2] College of Information Science and Engineering, Ningbo Univertisy, 315211，China

E-mail: [1]maoqinjiao@163.com , bqfeng@mail.xjtu.edu.cn, [2]panshanliang@nbu.edu.cn

ABSTRACT

In recommender systems, modeling user interest is a basic step to understand user's personal features. Traditional methods mostly just use the items that the target users navigated as their interests, which makes the inherent information unclear to the system and thus the recommendations are not intelligent enough. In this paper, we investigate the utility of topic model called LDA for the task of modeling user interests which are assumed as the latent variables behind user activities in the systems. By using such a probabilistic model on the generation of user profiles, the relationships amongst user、item and interest are constructed. Each user can be considered as a generation of the model. Based on this user interest model, we propose three item ranking methods for personalized recommendation. In order to get a better model for each algorithm, we use a simple automatic parameter turning way to select the model parameters. After carefully selecting the parameters, our methods can all receive encouraging results for recommender systems.

Keywords: *Topic Model, User Interest, Recommender System, Collaborative Filtering*

1. INTRODUCTION

When people are surfing on the internet, one of the first things they may do is to identify what kinds of links or web pages they are interested in, which means they are willing to click or navigate. We believe that there are such things called user interests which guide the user activities. As in the real world, people all keep distinct user interests which guide them to seek personalized information different from each other. A page they navigate may contain something related to their interests they want to get. Users keep the topic they follow in mind and this information plays a role in the assessment of whether an item is relevant to their interests. Since users on the web are suffering from information overload, catching user interests is very helpful to improve user experiences in nowadays web.

Recommender systems are becoming increasingly popular thanks to their utilities on providing people with recommendations of items they might be interested in and thus purchase or take a deep look at[1]. The systems learn users' preferences and recommend products they are expected to find from the large scale of all available goods. Therefore the objective of the system is to provide people with personalized experience to match their needs as the system is just designed only for the target user.

In natural language process, a topic model is a type of statistical generative model proposed firstly for analyzing latent abstract topics in a collection of documents. The most common model in use is called Latent Dirichlet Allocation(LDA) introduced by Blei[2], and it leads a new direction in this research. Afterwards, lots of similar topic models are proposed to deal with more complexity situations. A limitation of LDA is the inability to model topic correlation which stems from the use of Dirichlet distribution to model the variability among the topic proportions. A correlated topic model was proposed where the topic proportions exhibit correlation via the logistic normal distribution instead of Dirichlet distribution[3]. To capture topic evolvement in temporal data, how to integrate timestamps into topic models has been investigated. The dynamic topic model[4] simply divide documents into several subsets according to their timestamps and build topic modes for each subset and transformations between these models. The dynamic mixture model[5] assumes that the mixture of latent variables for all streams is dependent on the mixture of the previous timestamp. These models are all Markov chain-based models that put the Markov assumptions on the topic states transitions. There are also models that do not assume the Markovian dependence over time, for example, the topic over time[6]. In this model, timestamps are drawn from the same beta

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

distribution for topics. Though all of these models take the temporal data into consideration, timestamps are all connected to the documents, that is, one timestamp for each document since they are models for text analytics.

In fact, user interests can be regarded as topics to some extent. Topics can be used to present user interests very well. A web user profiling and clustering framework based on LDA-based topic modeling with an analogy to document analysis in which documents and words represent users and their actions was proposed by Hiroshi Fujimoto et al.[7] While most of the existing works are focus on information retrieving[8] or querying[9, 10], they can not easily be adapted directly on some of the recommender systems. In this paper, we present a user interests modeling method using LDA without considering the text of the items since lots of items in the web like movies or music can not be described well in texts. [8]Based on the interests we infer from this model, we propose three kinds of recommendation methods called pure-LDA, LDA-knn, and LDA-tran. Such a user interests modeling method is different from the traditional content-based interest modeling in that they are defined as mixture of items. It can be used quite exactly for recommender systems, and the utility for personalized recommendation will be discussed in our paper.

2. USERS, ITEMS AND INTERESTS

In a recommender system, a user is commonly described by the items they navigate or like. Items can be all kinds of web objects we may deal with, for example, web pages, movies, songs or goods. The system keeps user's activities on items, and takes advantage of them to figure out the preference on the remains ones they have not visited. A user may keep several interests in mind, and each interest has lots of items related to it in the system. A user usually only needs a tiny fraction of the items based on their interests on certain times instead of the all items under that interest all the time. So, a good recommender system is to provide users with the right item at the right time. Existing recommender systems are mostly aiming at providing users only the right item without considering the time factor. Somehow this hinders them from being a good recommender system.

Therefore, interests can be regarded as latent factors in recommender systems. Interests carried by the user determine which items the user wants to get. Unfortunately, it is hard for user to describe or provide what their interests are. Even though a user may tell what kind of things he/she interested in, making a recommendation is still difficult. The causes of this problem are manifold. First, the interests described by users themselves are always arbitrary. Therefore one interest may be described differently by different users, while different interests may be described as one thing. Second, even user interests are reliable to get, recommender system need to cnstruct the relationship between items and interests, and then decide which items are fit for? their interests amongst the large scale of candidates. Ignoring the time information, user-item dataset has the same form as word-document co occurrence matrix, with each user being a document, each item being a word, and a user visiting on an item being a document contains a word.

3. TOPIC MODEL FOR USER INTERESTS

In this paper, we build user interests using topic model called Latent Dirichlet Allocation(LDA). Since we are intending to model user interests, we will unify the concepts of topic and interest and use interest for both of them most of the time. Under the context of recommender system, we map the objects 'user, item, interest' we deal with to the objects 'document, word, topic' in the orignal topic model respectively. So, similarly, we can define the generative process about users and their item profiles. In the generative process, a multinomial distribution θ_j over interests is randomly sampled from the Dirichlet distribution with parameter α for each user u_j , then an interest z is sampled from the distribution θ_j , and an item is generated randomly from the interest distribution on items with parameter φ which is a sample of a Dirichlet distribution with paremeter β .

The graphical model representation for the model is given in Figure 1, and a brief notation about the symbols used in this paper is summarized in Table1.

There are lots of researches on extending the LDA model to model more complicated situations. While in this paper , we only dig into the utility of the topic model on recommendations.

Journal of Theoretical and Applied Information Technology
10th February 2013. Vol. 48 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

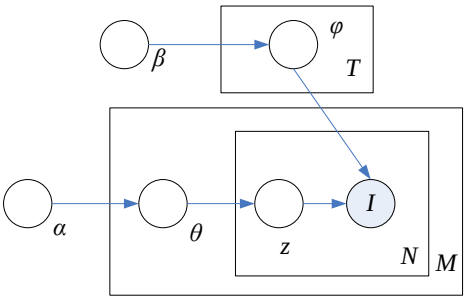

Figure1：Graphical Representation For LDA

Table1:List Of Notations Used In This Paper

Symbol	Description
T	Number of user interests
M	Number of users
N	Number of unique items
θ_i	The multinomial distribution of interests to user i
φ_z	The multinomial distribution of words to interest z
m_{dz}	The number of tokens in document d assigned to interest z
n_{zw}	The number of tokens of word w assigned to interest z
α, β	The parameters of the corresponding distributions

4. RECOMMENDATION ON USER INTERESTS MODEL

4.1. Ranking Items Based on User Profile

In recommender systems, users' activities are collected to construct user profiles. Based on users' past navigated items, we can infer their related interest distribution. Personalization that takes advantage of the interest distribution of the user can be derived as follows. For a user u_j with the interest distribution of θ_j , which is derived from the LDA-based model, we calculate the probability of the item I_i that the user may choose as follow:

$$P(I_i \mid u_j) = \sum_{t=1}^{T} P(z=t \mid u_j) \times P(I_i \mid z=t) = \sum_{t=1}^{T} \theta_{jt} \times \varphi_{ti}$$

It follows the generation process of the topic model on user profile, and it makes a simple assumption that the user's interests are depended on the past navigation information. Thus the items are ranked according to the probabilities and we call this method "pure-LDA" recommendation.

4.2. Ranking Items Based on User Similarity

The ranking method we proposed in 4.1 will suffer the problem of cold start. A new user with few items known by the system will make the topic distribution over fitted by the little information about the user. The user may start using the system for tiny times and has no clear or explicit intent except several clicks on some items. The information we have about the target user is insufficient. So, it is dogmatic to simply take these items as the user's preference. So, we propose another way of item ranking to do the recommendation. We learn the idea from collaborative intelligence that we find similar user to the target user based on the topic model and then make a recommendation based on the similar user. The most similar users to the target user are the ones who have the maximum conditional probability of the item set of the target user, given the candidate users. Given a target user u_j who has visited the items I_{u_j} , we calculate the probability of generating I_{u_j} under the condition of the existing known users u_i by

$$P(I_{u_j} \mid u_i) = \prod_{I \in I_{u_j}} P(I \mid u_i)$$

$$= \prod_{I \in I_{u_j}} \sum_{t=1}^{T} P(z=t \mid u_j) \times P(I \mid z=t)$$

$$= \prod_{I \in I_{u_j}} \sum_{t=1}^{T} \theta_{jt} \times \varphi_{ti}$$

We select a user set S_j , $|S_j| = K$ for user u_j in which the users $u \in S_j$ have the largest $P(I_{u_j} \mid u)$. Then we calculate the preference on unknown items I_i for user u_j according to $P(I_i \mid u_j) = \sum_{u_i \in S_j} P(I_i \mid u_i)$.

The probability $P(I_i \mid u_i)$ is calculated in the same way with "pure-LDA" in 4.1. We call the recommendation based on this ranking method "LDA-KNN".

4.3. Ranking Items Based on Item Transition

Based on the topic model LDA we build, there is additional information we can get, which we call

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

item transition. Transition is one relationship between items, indicates the probability that one item will be visited after another. Thought the transition is calculated by the usage data, it inherits similarities or associations between items. Transitions from item I_1 to I_2 can be expressed by a conditional distribution of accessing I_2 when having the clue of I_1. In our generative model, items are generated randomly by the topic distribution. It seems like each item is generated independently and has no connection to other ones. While in fact, we can get the associative relationships on items according to the model we infer. Assuming that each item is only decided by one interest, we can get the postorir distribution $P(z = j \mid I_1)$ from the model. Then the conditional distribution of I_2 under I_1 can be calculated by:

$$P(I_2 \mid I_1) = \sum_{j=1}^{T} P(I_2 \mid z = j) P(z = j \mid I_1)$$

Here, the posterior distribution is calculated by:

$$P(z = j \mid I_1) = \frac{P(I_1, z = j)}{\sum_z P(I_1, z)}$$

Based on the item relationships we get, we calculate the preference on unknown items for user u_j by summing up all the probabilities that can be transformed from the items that the user has ever visited:

$$P(I \mid u_j) = \sum_{I_i \in I_{u_j}} P(I \mid I_i)$$

We call the recommendation based on this ranking method "LDA-tran".

4.4. Parameter Tuning Method

In each algorithm we proposed so far, the recommendation results are sensitive to the parameters α, β, T we use to estimate the model. It's hard to find global optimum parameters. We need to tune the parameters in order to maximize the recommendation precision S@K. We use a simple random searching method called Automatic Parameter Tuner (APT) to find the relative better solution to the model.

The basic idea is to randomly change one of these parameters, check if the result gets better under the new parameter value, and then decide if keep the value or not. In detail, this works as follows:

Randomly select a number i from $\{1, 2, 3\}$ and draw a new parameter value for the i th parameter. If the new parameter makes the result better than the old one, we assign the new value to the parameter. Loop this process until the result is good enough.

When dealing with LDA-knn, we add the parameter K into the tuning method, which makes i selected from $\{1, 2, 3, 4\}$.

5. EXPERIMENTS

5.1. Evaluation Protocol

In order to examine the effectiveness of our models, we conducted an experiment on the real world dataset call MovieLens. The data set was collected from the GroupLens research site. It contains 943 users, 1682 items and a total of 100,000 rating data. Each user has no less than 20 ratings. The data is randomly divided into a training set and a test set with exactly 10 ratings per user in the test set. After the division, each user has at least ten ratings as his profile for training. The experiments were implemented using a modified version of the "Matlab Topic Modeling Toolbox 1.4", provided by Mark Steyvers and Tom Griffiths.

5.2. Evaluation Metrics

For the performance measures, we use the metric $s@k$ which is defined as:

$$s@k = \frac{1}{Q} \sum_{i}^{Q} 1(r(u_i, I_i) \le k)$$

Here, Q is the number of user-item pairs we used in the test set, $r(u_i, I_i)$ is the ranking of item I_i for user u_i and $1()$ is an indicator function which returns 1 when its argument is true and 0 otherwise. $r(u_i, I_i)$ is corresponding the descent order of $P(I_i \mid u_j)$.

The metric $s@k$ only considers the number of items that returned in the recommended lists without considering the order of the items. For a more exhaustive analysis, we employ another metric called the mean reciprocal rank (MRR):

$$MRR = \frac{1}{Q} \sum_{i}^{Q} \frac{1}{r(u_i, I_i)}$$

Journal of Theoretical and Applied Information Technology
10th February 2013. Vol. 48 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We are hoping the items in the test set with smaller $r(u_i, I_i)$ that can be recommended in the head of the lists, such that the bigger $s@k$ or MRR 's value is, the better the algorithms performance.

5.3. Experimental Results

In this section, we focus on how the evaluation measures evolve with the parameters. Based on the two metrics, our experiments are done in either higher $s@k$ targeted or higher MRR targeted. The two targeted APTs receive results a little different. In order to tune parameters to have a better result for recommendation, we iteratively choosing a new value for the randomly selected parameter to see if it improves $s@k$ of the recommendation or not to decide if parameter will be kept or replaced by the new value. By doing this with many different initial values and keep the iteration until the metric stop changing for a relative long time, we record the best results. And this process is done similar with the objective of higher MRR .

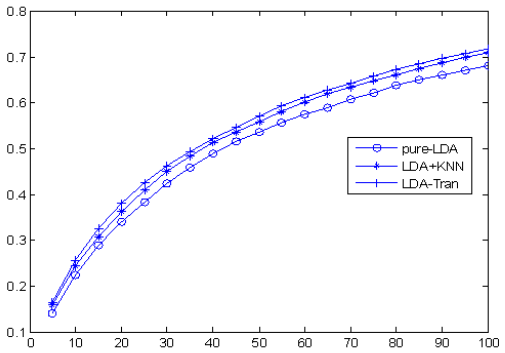

Figure 2 $s@k$ Of Top-K Recommendation Using LDA With Higher

In Figure 1, we present the results of recommendation precision $s@k$ based on the topic models with higher $s@k$ targeted and the corresponding parameters and MRR s are listed in Table 2. As we can see that the performance of Pure-LDA is relative worse that the other two methods, and LDA-tran is the best one of the three. By choosing parameters with higher MRR s, we get the corresponding parameters and MRR s listed in Table 3 and $s@k$ in Figure 3.

Table 2: Corresponding Parameters And MRR S For Recommendation Using LDA With Higher $s@k$ Targeted.

Method	α	β	T	K	MRR
PureLDA	1.0707	0.0145	15	--	0.1060
LDAKNN	0.1667	0.0096	70	14	0.1121
LDAtran	0.1667	0.0118	96	--	0.1148

Table 3: Corresponding Parameters And MRR S For Recommendation Using LDA With Higher MRR Targeted.

Method	α	β	T	K	MRR
PureLDA	0.8922	0.0142	22	--	0.1085
LDAKNN	0.1667	0.0096	70	14	0.1164
LDAtran	0.1667	0.0085	70	--	0.1187

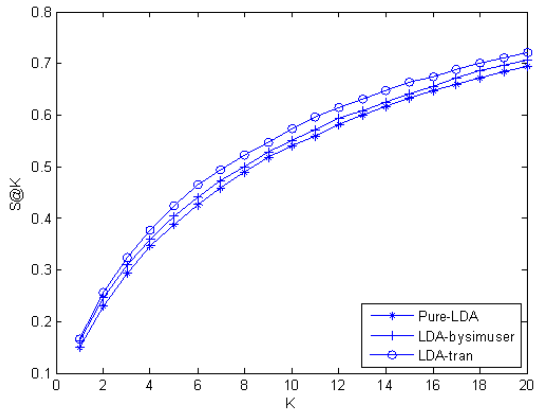

Figure 3: $s@k$ Of Top-K Recommendation Using LDA With Higher MRR

Besides, we investigate the impact of user profile size on the algorithms' precision. We divided the users by their item numbers in training set into [10,30), [30,60) , [60,100) , [110,200) , [200,300), , [300,717] and mark the corresponding results as Profile-30, Profile-60, Profile-100, Profile-200, Profile-300, Profile>300. Figure 4 shows the results of the three algorithms on different subsets. Form the results, it is interesting to notice that all of the three algorithms perform better when the profiles are relative smaller, which means that the items we want to predict are decided by few items and the precisions will be declined if we take too much items as the clue to the specific items. So, in real

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

systems, setting a time window to limit the user profile may be helpful in making the prediction.

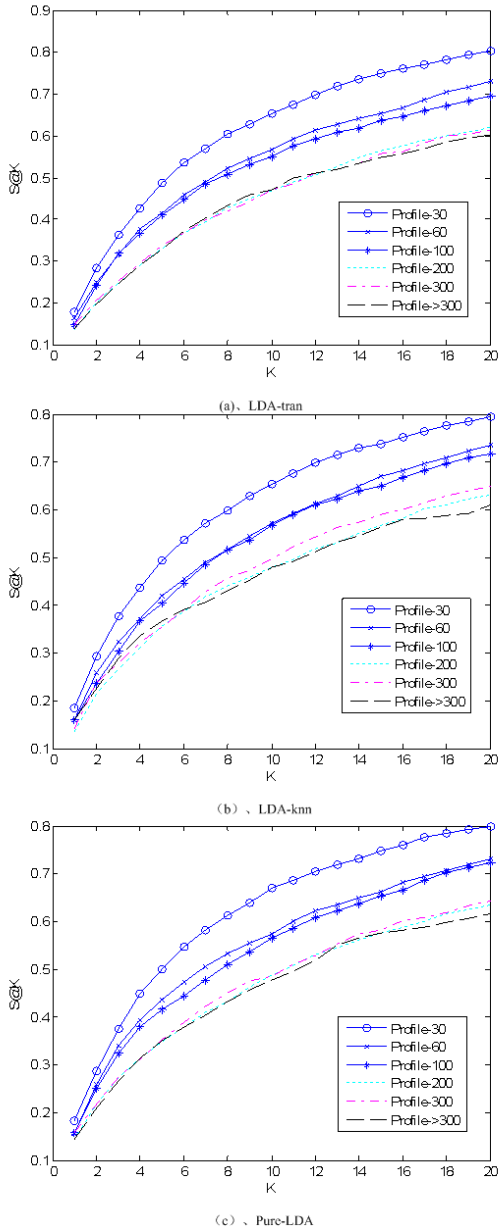

Figure 4: $s@k$ Of Profile-30, Profile-60, Profile-100, Profile-200, Profile-300, Profile>300.

6. CONCLUSION

In this paper, we view user interests as latent variables hidden in the recommender system based on topic model. User interests are represented by multinomial distributions on items, and users are modeled as multinomial distributions on interests. Using Gibbs sampling methods, user interests are inferred from a generative model according the training data. We examine the model effectiveness in recommender systems by three ranking methods on the model and discuss the results. Comparing to the previous works, topic model used in this paper can build the relationships on user, item and user interest which are different from the traditional way and quite helpful to construct algorithms to make recommendations. In this paper, we propose three kinds of methods to rank items for users. The experiments on the data of recommender system show the effectiveness of these algorithms qualitatively and quantitatively. Though our model seems quite effective in recommendations, our methods still have some shortages. In our methods, we made an assumption that user interest keep still during the whole period, thus the model does not consider the temporal information.

As in real world, user's personal preference in their interests also changes during the time, that is, the user profiles are changing too. This problem may be the work we will deal with in the following work.

ACKNOWLEDGMENTS:

This work is supported by the National Natural Science Foundation of China (No.61202181, No.61173040) , Ningbo Natural Science Foundation (No.2012A610066) , Zhejiang Provincial Natural Science Foundation (No.LY12F02020).

REFRENCES:

[1] C. N. Ziegler, G. Lausen and L. Schmidt-Thieme, "Taxonomy-driven computation of product recommendations," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 406--415.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993--1022, 2003.

[3] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, vol. 1, pp. 17--35, 2007.

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning,ICML '06*, New York, NY, USA, 2006, pp. 113--120.

[5] X. Wei, J. Sun and X. Wang, "Dynamic mixture models for multiple time series," in *Proceedings of the 20th international joint conference on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Artifical intelligence,IJCAI'07, San Francisco, CA, USA, 2007, pp. 2909--2914.

[6] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2006, pp. 424--433.

[7] H. Fujimoto, M. Etoh, A. Kinno, and Y. Akinaga, "Web user profiling on proxy logs and its evaluation in personalization," *Web Technologies and Applications*, pp. 107--118, 2011.

[8] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, "Characterizing web content, user interests, and search behavior by reading level and topic," in *WSDM '12*, New York, NY, USA, 2012, pp. 213--222.

[9] L. Li, G. Xu, Z. Yang, P. Dolog, Y. Zhang, and M. Kitsuregawa, "An efficient approach to suggesting topically related web queries using hidden topic model," *World Wide Web*, pp. 1-25, 2012.

[10] M. J. Carman, F. Crestani, M. Harvey, and M. Baillie, "Towards query log based personalization using topic models," in *Proceedings of the 19th ACM international conference on Information and knowledge management,CIKM '10*, New York, NY, USA, 2010, pp. 1849--1852.