International Journal of Innovative Computing, Information and Control
Volume **12**, Number **6**, December **2016**

ICIC International ©2016 ISSN 1349-4198
pp. **2099–2110**

MINING LEARNERS' TOPIC INTERESTS IN COURSE REVIEWS BASED ON LIKE-LDA MODEL

Xian Peng, Sanya Liu*, Zhi Liu*, Wenbin Gan and Jianwen Sun

National Engineering Research Center for E-Learning
Central China Normal University
No. 152, Luoyu Road, Wuhan 430079, P. R. China
*Corresponding authors: { lsy5918; liuzhi8673 }@gmail.com

Received May 2016; revised September 2016

Abstract. *The diversity and volume of textual data in the course reviews area are overwhelming. These data offer faculties with a chance to capture topic interests of learners and further make relevant recommendations for them. "Like" or "+1" considered as a specific interactive means among learners has been commonly applied to enriching the interactivity and flexibility of online community. This information can reflect their supporting, appreciation for the textual contents and topical interests. In view of this situation, the topic model by incorporating the behavioral feature "like", namely Like-Latent Dirichlet Allocation (Like-LDA), is proposed to detect the latent topic interests in course reviews. The experimental results on the real-life dataset show that, Like-LDA can gain better performance in extracting the new hidden topics, higher accuracy of topic detection and words coherency in each topic than LDA.*

Keywords: Topic interests, Topic modeling, LDA, Like, Course reviews

1. **Introduction.** Massive Online Open Course (MOOC) is leading the rapid development of global higher education, and has become a new form of learning. Unlike the traditional face-to-face teaching, it is hard to instantly track learners' explicit behaviors and potential inner psychological states, owing to the separation nature of space and time of distance teaching. This has brought great challenges to further realize the personalized service supports. For them, without considering the influence of interest in the online teaching, it will lead to the cognitive load, triggering learning boredom, anxiety and other adverse psychological tendency, and even affecting the course completion rates [1-3]. Nowadays, interactive technologies in the application of online learning environments have promoted the emergence of large-scale unstructured data, e.g., course reviews, discussions, questions, logs, sound, images, and audio. In particular, course reviews are not only treated as an important source of learner-generated text data, but also are thought to be the primary way of communication for them to express their interesting topics about courses, teachers and designs of function in MOOC.

With the growing volume of learner-generated reviews, the large-scale online education is coming to screening period from conventional accumulation period of contents. Learners are more willing to share their personal learning experiences with others. However, it will be time-consuming for manually analyzing all online reviews to interact with learners interested in the same topic contents, to identify the suitability of courses for them, and to make some immediate feedbacks for these learners met with learning problems. Therefore, it is imperative to automatically mine the useful topical information within these subjective contents. To detect potential and hidden topics from a large unstructured collection of documents in an automatic way, Latent Dirichlet Allocation (LDA) [4] model first proposed by Blei et al. was considered as a typical topic modeling method, which had been widely applied in various domains in recent years, especially in business intelligence field [5,6]. Titov and McDonald [7] put forward a novel framework for extracting multi-grain aspects of objects from online user reviews by extending standard LDA model. Moghaddam and Ester [8] introduced an Interdependent Latent Dirichlet Allocation (ILDA) model to extract a rated aspect summary of product reviews. Li et al. [9] presented the topic distribution of the news collection as a topic vector based on the LDA model, promoting the intelligent orientation of advertising marketing and personalized news customized service. Jiang and Sha [10] adopted semantic enrichment method to discover user interests, and used the topic hierarchy tree model, in the Twitter social network, to capture user interests change over time. However, in the educational field, there are few studies automatically discerning learners' topical interests by using topic modeling methods. Daud [11] demonstrated a temporal topic modeling approach called Temporal-Author-Topic (TAT) to capture the change of researchers' study interests over time, which can simultaneously deal with the exchangeability of topics problem. Ramesh et al. [12] proposed a seeded LDA model to uncover useful thematic information in online discussion forums, aiming to help predict student survival. Building on the above proposed seeded LDA method, Ramesh et al. [13] developed a weakly supervised joint model for aspects and sentiment in online courses using the Hinge-Loss Markov Random Field (HL-MRF) probabilistic modeling framework.

Although the goal of the above existing studies has mostly focused on how to analyze and capture the textual semantic information, it is inefficient owning to the lack of explicit behavioral characteristics associated with the textual content, e.g., "like". Moreover, understanding the way users interact with the text will help to build users' topic interest profiles. Recent advances in mining users' topic interests allow us to combine behavioral features with textual contents [14-16]. These behavioral features are derived from the context in which the text is generated and interacted with, such as social networks like Twitter and Facebook which provide users with several kinds of common ways of interaction, e.g., "post", "retransmission", "reply" and "mention". For example, Zhao et al. [17] employed matrix factorization techniques involving various interactive behaviors to model users' topic interest profiles, acquiring better results than baseline methods.

Intuitively, different behavior actions in online reviews indicate learners' attention on diverse textual topic information. For example, an online learner might post comments about course contents, video production and the function of online learning platform. However, he might typically click "like" button on posts about traveling, reading, shopping, etc., preferring to keep his interests somewhat private. While these reviews learners post reveal their interesting topics, different behavior actions could also provide insights into online learners' topic interest profiles. Therefore, in order to more exactly construct the learners' interest profiles, it will be essential to integrate this special behavioral feature into the procedure of mining their topic interests. In addition, our study is based on the assumption that learners, to some extent, are interested in the comment contents they "like". By this idea, a novel model, called Like Latent Dirichlet Allocation (Like-LDA) model, is developed. To our knowledge, this is the first work to extend standard LDA model by embedding the specific behavior feature in the application of online learning setting.

The remainder of this paper is organized as follows. Section 2 introduces the basic notion of standard LDA. Section 3 depicts the improved model Like-LDA, and illustrates the main steps of the model generation and topic detection algorithm. Section 4 displays the whole experimental process and results in detail. Finally, we make a conclusion for this study in Section 5.

2. **Probability Topic Model.**

2.1. **Probability topic model.**

Probability topic model [18] is an unsupervised machine learning method, which is mainly applied to automatically identifying the potential thematic information in large-scale electronic archives or corpus. Obviously, documents can be easily observed, while the thematic structure, e.g., document-topic distribution and topic-word distribution, is hidden. Therefore, the central issue of topic modeling is to observe how to generate the sequence of words in a corpus and infer the latent structure of textual contents. Topic model was originated in the Latent Semantic Indexing (LSI) [19] presented by Deerwester et al., which laid a solid foundation for the development of model. In 1999, on the basis of LSI, Hofmann [20] proposed Probabilistic Latent Semantic Indexing (PLSI) that was considered as the topical model in true sense. In 2003, Blei et al. [4] introduced Latent Dirichlet Allocation (LDA) by extending the PLSI.

2.2. **Introduction of LDA model.**

LDA is a hierarchical Bayesian model that contains words, topics, documents in three levels, as shown in Figure 1. The basic notion is that each document is composed of several topics, and each topic is represented by some words. We can simply understand that the generative process of LDA model is a joint probability distribution including the observed and hidden random variables. By using the joint distribution, we conduct data analysis to calculate the conditional distribution (posterior distribution) of the hidden variables given the observed variables. For LDA, the observed variables are the words of the documents $w_{d,n}$, and hidden variables are the thematic structure in the documents z_k. The set of notations of the standard LDA model is described in Table 1.

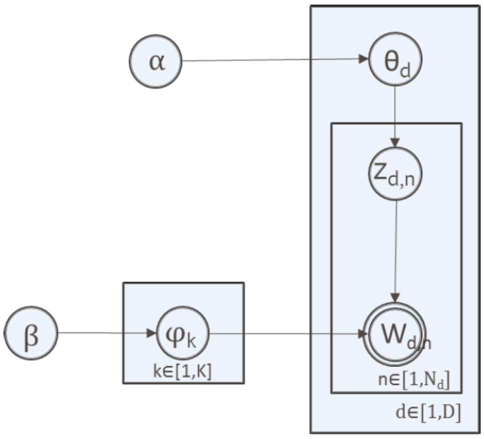

FIGURE 1. The standard LDA model

LDA model employs the Dirichlet distribution as a prior distribution of multiple distribution in probability topic model. In Figure 1, α and β are respectively the prior parameters of document-topic probability distribution θ_d and the topic-word probability distribution φ_k; each node represents a random variable, and arrows between nodes imply certain dependencies; for example, a random variable $w_{d,n}$ is dependent on the θ_d and φ_k. As to the generative procedure of LDA, it can be divided into two-stage steps as follows.

First, select a distribution over topics.

Second, for each word in the document:

a. select a topic from the document-topic distribution;

b. select a word from the topic-word distribution.

Here, we just simply describe each document in the collection that is how to be generated based on the core ideas of LDA modeling. In the next Section 3.1, we will present the detailed generative process of our proposed model Like-LDA.

Table 1. Descriptions of notations of Like-LDA

Notations	**Description**
α	Dirichlet prior knowledge of learner-topic probability distribution
β	Dirichlet prior knowledge of topic-word probability distribution
β_l	Dirichlet prior knowledge of topic-word probability distribution
θ_d	Distribution of topics for the learners' original reviews
φ_k	Distribution of words for the kth topics from comments learners post
$\varphi_{k,l}$	Distribution of words for the kth topics from comments learners post and "like"
$z_{d,n}$	Topic of the nth word in the review d
$w_{d,n}$	The nth word in the review d
s	Control variable 0 or 1
V	The vocabulary size
K	The total number of topics
D	The total number of course reviews
U	The total number of learners

3. **Inference on Like-LDA.**

3.1. **Like-LDA model algorithm description.** It is a common phenomenon that learners tend to post comments as the primary participation way in learning community. Nowadays, in order to strengthen and enrich the flexibility of online community, "like" considered as a specific interactive means has caused much attention [21,22]. Intuitively, the explicit behavioral characteristic of online learners, to some extent, can reflect their support, appreciation for the textual contents they like and their topic interests. The following is a concrete example of one comment the learner named mike123 liked. For example, "Mrs Xiao was a financial analysis teacher, who tended to give a lesson in a way of simple words. With the aid of the representation of some cartoons, it was easier to explain and understand contents she taught in online class. And the class was vivid, not

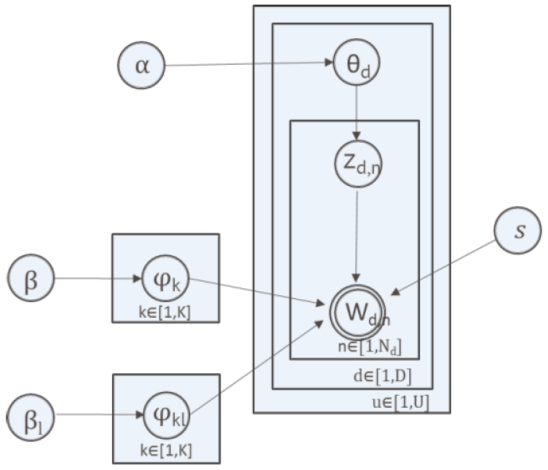

Figure 2. The improved Like-LDA model

boring. Be great!" From this comment, we can find that mike123 shows some interests in the style of teaching and the course contents.

As our main contribution, based on the LDA model, Like-LDA model is proposed combining with the feature of "like" in the online reviews, which not only contains original comments by individual learner, but also comprises the others' textual contents one likes. As shown in Figure 2, it is a probabilistic graph model; s is a control variable and equals 0 or 1; and $\varphi_{k,l}$ denotes the distribution over topics from the collection of reviews individual learner posted and liked. The set of notations of the Like-LDA model is explained in Table 1.

Now, the Like-LDA model will be presented in detail. First, we assume that there are u online learners participating in posting course reviews. Each learner may post comments or "like" others' reviews, which can be defined as $u = \{r_1, r_2, \ldots, r_d\}$ $(1 \leq d \leq D)$. Second, we make an assumption that there are K assigned topics, where each topic has a multinomial word distribution $\varphi_{k,l}$. And $\varphi_{k,l}$ can be described as $z_k = \{w_1, w_2, \ldots, w_n\}$. In this study, note that each document is corresponding to the level of course review. Each review can be exhibited with multiple topics. Therefore, distribution of topics for the learners' original reviews is θ_d, which can be represented as $u = \{z_1, z_2, \ldots, z_k\}$. For the words w $(1 \leq w \leq V)$ in each review, they are sampled from the φ_k or $\varphi_{k,l}$. When one word w_n is randomly sampled, it belongs to a specific topic z_k and depends on the value of $s \in \{0, 1\}$, which denotes a control variable. If s equals 0, the word is sampled from φ_k; if s equals 1, the word is sampled from $\varphi_{k,l}$. At last, we assume that θ_d, φ_k and $\varphi_{k,l}$ have Dirichlet prior knowledge α, β and β_l respectively. The elaborate generative process of Like-LDA model is depicted in the following six-stage process.

```
 Input:  The learners' reviews collection $U = \{c_1, c_2, \ldots, c_d\}$ $(1 \leq d \leq D)$
         The number of iteration of Like-LDA: $N$
         The number of random sampling topics: $K$
         The value of control variable $s$
         Dirichlet priors: $\alpha$, $\beta$, $\beta_l$
Output:  The learner-topic matrix $\theta$ and the topic-word matrix $\varphi$
Step 1:  Initialize the model and the related parameters $K$, $s$, $\alpha$, $\beta$, $\beta_l$.
Step 2:  For each topic $z_k$, $k = 1, \ldots, K$
         If the control variable $s$ equals 0, randomly perform multiple samples from word
         probability in topics: $\varphi_k \sim Dir(\beta)$.
         If the control variable $s$ equals 1, randomly perform multiple samples from word
         probability in topics: $\varphi_{k,l} \sim Dir(\beta_l)$.
Step 3:  For learners' each review $d_i$, $i = 1, \ldots, D$
         Randomly conduct multiple samples from the topic proportions in reviews: $\Theta_d \sim$
         $Dir(\alpha)$.
Step 4:  For each word of review $w_{d,n}$, $n = 1, \ldots, N$
     a)  Randomly process multiple samples from the distribution over topics in reviews,
         and get the topic $z_{d,n}$ of $w_{d,n}$, $z_{d,n} \sim Multi(\theta_d)$.
     b)  Randomly process multiple samples from the distribution over words in topics
         $z_{d,n}$, and get the word $w_{d,n}$, $w_{d,n} \sim Multi(\varphi_{z_{d,n}})$.
Step 5:  Conduct Gibbs Sampling method to assign new topic label $z_k$ for each word $w_{d,n}$
         and update the model parameters.
Step 6:  Repeat until the optimization of the model, and compute the proportion of topics
         learners tend to discuss and comment $\theta$ and the proportion of words belonging
         to a special topic $\varphi$.
```

3.2. **Model inference.** Given a collection of course reviews, $w_{d,n}$ can be observed in the known variables; α and β are prior parameters configured by the past experience. We need to infer and estimate the rest hidden variables by the observed variables. According to the dependencies in the Like-LDA model, the joint distribution covering all random variables is as follows.

$$\begin{aligned} p\left(w_{d,n}, z_{d,n}, \theta_d, \varphi_{k,l} | \alpha, \beta_l\right) &= \prod_{d=1}^{N_d} p\left(w_{d,n} | \varphi_{z_{d,n}}\right) p\left(z_{d,n} | \theta_d\right) p\left(\theta_d | \alpha\right) p(\varphi_{k,l} | \beta_l) \\ &= \prod_{z=1}^{K} \frac{\Delta(n_z + \beta_l)}{\Delta \beta_l} \cdot \prod_{d=1}^{D} \frac{\Delta(n_d + \alpha)}{\Delta \alpha} \end{aligned} \tag{1}$$

In this way, θ_d and $\varphi_{k,l}$ are the ultimate unknown parameters we should solve. However, it is difficult to solve them accurately. Therefore, Griffiths et al. [23] proposed a faster and easier algorithm to figure out the approximate inference problem called Gibbs Sampling, which is a special case of the Markov Chain Monte Carlo method. The general idea is as follows.

Firstly, randomly assign a topic for each word in each review, and constitute the initial state of Markov Chain.

Secondly, according to the distribution of all other topics z_{-i}, adopt the following Formula (2) to estimate the current word w_i in the proportion of each topic, and sample a new topic t to w_i. After all words in each review are iterated, the Markov Chain steps into the next state. For Formula (2), without considering the current item w_i in the course review d_i, $n_{d,-i}^{(k)}$ is the total number of words belonging to the topic t; without considering the current item w_i in the corpus, $n_{k,-i}^{(t)}$ is the total number of words belonging to the topic.

$$p\left(z_i = t | \omega_i, \alpha, \beta_l, Z_{-i}\right) \propto \frac{\alpha + n_{d,-i}^{(k)}}{\sum\limits_{k=1}^{K} \left(\alpha + n_{d,-i}^{(k)}\right)} \cdot \frac{\beta_l + n_{k,-i}^{(t)}}{\sum\limits_{v=1}^{V} \left(\beta_l + n_{k,-i}^{(t)}\right)} \tag{2}$$

Thirdly, repeat Step 2 until the steady state the Markov Chain comes to. Then we can estimate the learner-topic matrix θ and the topic-word matrix φ by analyzing the posterior probability of the proportion of topics in reviews and the proportion of words in topics.

$$\theta_{d,k} = \frac{\alpha_k + n_{d,-i}^{(k)}}{\sum\limits_{k=1}^{K} \left(\alpha_k + n_{d,-i}^{(k)}\right)} \tag{3}$$

$$\varphi_{k,t} = \frac{\beta_{lt} + n_{k,-i}^{(t)}}{\sum\limits_{v=1}^{V} \left(\beta_{lt} + n_{k,-i}^{(t)}\right)} \tag{4}$$

In this part, with the aid of Formulas (1) to (4), the hidden variables θ and φ from the Like-LDA are deduced and computed. Finally, through logical reasoning, we can detect the distribution of the hidden topic interests of learners from the learner-topic matrix θ. Moreover, we can discover more fine-grained contents learners focus on from the topic-word matrix φ, which provides a decision-making basis to further implement intelligent tutor.

4. **Experiment.** In this section, we introduce and carry out the whole process of experiments. Besides, by applying our improved Like-LDA model on the review collections, the experimental results are presented in detail.

4.1. **Data sources.** The textual data sets in this study are captured by adopting the method of web crawler from the online reviews module of the mooc.guokr.com. It is one of the most publicly influential MOOC learning platforms in China. There are a large number of excellent curriculums sourcing from Couresa, Edx, Udacity, etc. Eventually, we got a total of 6163 comments from the 50 highest rated courses from December 2013 to December 2015, e.g., language, literature, science and engineering, management, and finance. In addition, we also set up a mapping relationship combining each learner with comments he/she posted and liked for further analysis.

4.2. **Data preprocessing.** In order to accurately parse textual data for our study, it is essential to perform text preprocessing. First of all, we need to divide each review into words by making use of the Chinese word segmentation system of Chinese Academy of Sciences ICTCLAS [24] and simply retain adjectives, adverbs, verbs and nouns. Then, we remove the stop words, noise words, low-frequency words and symbol, etc. Because of the expression of non-sense and non-standard cyberword from online learners, the user dictionary is established in this study to constrain these "special words", e.g., "菜鸟/newbie", "学神/super scholar", "果壳/guokr", "先修/prior knowledge".

4.3. **Experimental results.** By observing massive experimental data, we find some interesting phenomena. For example, some learners click "like" button on others' reviews many times, but never post; some learners only post reviews, but not like; and some learners almost like their own reviews, etc. In this case, we randomly select 10 learners who post and like a certain number of reviews, a total about 300 comments. They are more appropriate as our research object. The experimental data is divided into two groups: one set of the original comments they post, and the other set of comments they post and like. Then the experiment is carried out employing the LDA and the Like-LDA model respectively. The total iteration times of each model is processed 100. Starting from 60, we begin to save the model and record θ and φ. After every 5 times, the parameters of each model are updated and output. Based on the previous researches [4,7,8,12], β is generally set as 0.1. In this study, two evaluation indexes called similarity and entropy are utilized to quantitatively validate the effectiveness of model in Section 4.3.4. Through repeatedly adjusting the initial parameters, we observe that when K and α are set as 20 and 0.2 respectively, the values of two evaluation indexes are smaller. In other words, the model will be more stable and can gain better performance in experimental results.

4.3.1. *Results of Like-LDA.* Table 2 represents five significant topics learners are more likely to take part in discussions, and the top 10 words with the largest probabilities within each topic are listed as follows.

As shown in Table 2, it can be found that the topic1 is related to the feelings and experiences of learners about the MOOC courses. The topic15 is about the literature course named "A Dream in Red Mansions", one of the four greatest ancient Chinese novels. The topic18 involves the discussion of architecture. It is notable that compared with the standard LDA, our model can detect the new latent topic4 and topic7. The topic4 is more likely to discuss the basic knowledge of the introduction and planning of marketing. The topic7 refers to one historical course called "Records of the Grand Historian", one of the China's famous ancient history. From Table 3, we can see the most representative topical distribution of individual learner according to the topical probabilities. In this paper, topic probability above 0.06 can be regarded as one of the learners' interests. Thus, combined with the topic-word matrix, it is definitely intuitive to discover their interesting topics and detailed items. To sum up, our model can not only incorporate the distinct behavioral feature "like" in constructing learners' topic interest profiles, but

Table 2. The results of topic-word probabilistic matrix of Like-LDA model

Topic1 (0.056)	Topic4 (0.045)	Topic7 (0.042)	Topic15 (0.104)	Topic18 (0.071)
Words Prob.	Words Prob.	Words Prob.	Words Prob.	Words Prob.
收获/gain (0.020)	顾客/customer (0.022)	看/look (0.043)	观点/idea (0.044)	建筑/building (0.039)
学习/study (0.020)	中心/central (0.022)	史记/Shiji (0.0360)	听/listen (0.015)	文化/culture (0.037)
总结/summary (0.020)	广/wide (0.015)	作业/homework (0.029)	红楼梦/Honglou (0.015)	学习/study (0.022)
时间/time (0.013)	案例/case (0.015)	史为镜/history for mirror (0.015)	同意/admit (0.015)	视频/video (0.026)
评论/comment (0.013)	品牌/brand (0.015)	条件/condition (0.015)	言情/erotica (0.015)	好课/ good lecture (0.026)
慕课/MOOC (0.013)	营销/marketing (0.015)	圣人/sage (0.015)	主观题/ subjective (0.015)	修/repair (0.026)
情况/situation (0.013)	同学/classmates (0.008)	君子/gentleman (0.015)	形式/form (0.008)	收获/gain (0.026)
难/difficult (0.013)	听懂/ understand (0.008)	了解/know (0.007)	一般/general (0.008)	推荐/ recommend (0.026)
批评/criticize (0.013)	空洞/boring (0.008)	读/read (0.007)	好/good (0.008)	雄伟/grand (0.014)
思考/reflect on (0.013)	导读/ introduction (0.008)	脉络/vein (0.007)	字/word (0.008)	负担/burden (0.014)

Table 3. The results of learner-topic probabilistic matrix of Like-LDA model

Learner	Topic	Probability	Learner	Topic	Probability	Learner	Topic	Probability
Susan _4869	15	0.104	*Fang Cheng*	2	0.110	*Watter fall*	19	0.107
	12	0.096		4	0.098		14	0.091
	11	0.081		15	0.071		15	0.089
	10	0.071		10	0.071		3	0.084
	3	0.064		3	0.0569		11	0.080

also qualitatively uncover more informative topics and fine-grained aspects. Moreover, these information can provide faculties with a chance to make better recommendations for learners.

4.3.2. *Comparison of the words probability in one topic between LDA and Like-LDA.* In order to validate the extraction effect of the distribution of words in topic among two models, the experimental result is presented in Table 4.

As can be seen from Table 4, this topic is about the discussion of architecture. Furthermore, compared with LDA, by using Like-LDA, the overall words probabilities about this topic are slightly higher, and the distribution of words is more relevant to it.

TABLE 4. Comparison of the words probability in one topic between LDA and Like-LDA

The distribution of words in LDA	Prob.	The distribution of words in Like-LDA	Prob.
推荐/recommend	0.0380	建筑/building	0.0390
建筑/building	0.0380	文化/culture	0.0372
好课/good lecture	0.0257	视频/video	0.0260
修/repair	0.0257	好课/good lecture	0.0260
视频/video	0.0257	修/repair	0.0260
层面/level	0.0162	收获/gain	0.0260
学期/semester	0.0134	推荐/recommend	0.0260
感受/feeling	0.0134	雄伟/grand	0.0149

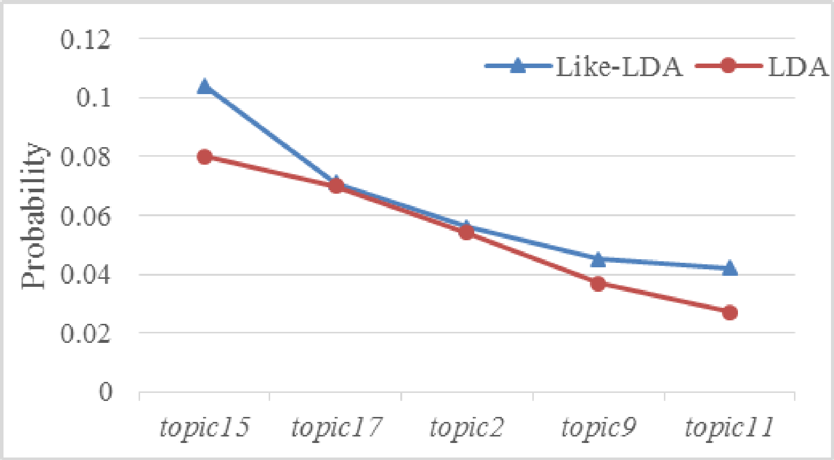

FIGURE 3. Comparison of different topical probabilities at the same learners

4.3.3. *Comparison of different topical probabilities at the same learners.* Taking the Susan_4869 as an example, the result is shown in Figure 3, and the topical probabilities operated by the Like-LDA are slightly higher than the LDA, which means that the topic interests of learners inferenced by the proposed model are better.

4.3.4. *Comparison of the model effectiveness.* To further validate the effectiveness of the model, two evaluation indexes are utilized to quantitatively explain the model. The smaller their values are, the better performance of the model is.

Similarity: It is a critical factor to measure the degree of similarity between topics.

$$Sim(z_i, z_j) = \frac{\sum_{w=1}^{V} \varphi_{iw}\varphi_{jw}}{\sqrt[2]{\left(\sum_{w=1}^{V} \varphi_{iw}^2\right)\left(\sum_{w=1}^{V} \varphi_{jw}^2\right)}} \tag{5}$$

Entropy: Topic entropy is the polymerization degree of word sequence which indicates the consistence and coherence of the internal information of topic.

$$Entropy = \frac{1}{K}\sum_{e=1}^{2}\sum_{k=1}^{K}\sum_{w=1}^{V}(-\varphi_{ewk}\log\varphi_{ewk}) \tag{6}$$

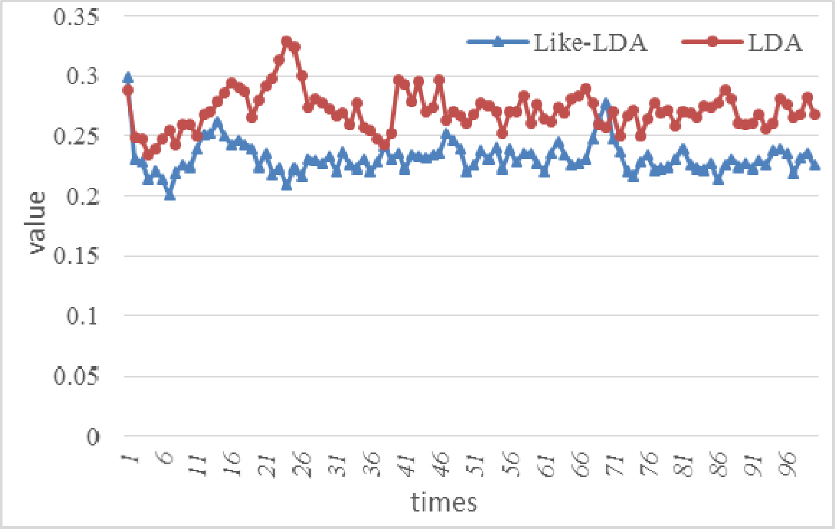

Figure 4. The comparison of similarity

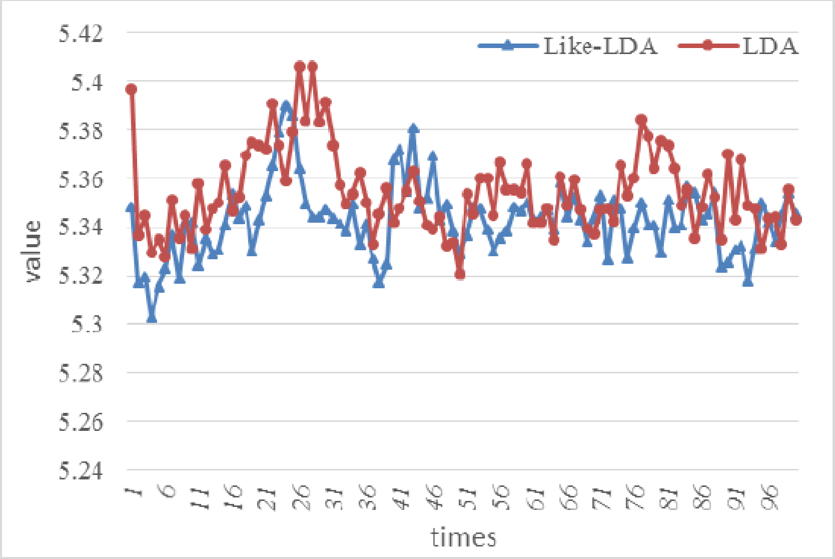

Figure 5. The comparison of entropy

As is presented in Figures 4 and 5, with the increase numbers of iterations, the evaluation value of our proposed model is generally a little bit better than LDA. It suggests that the effectiveness and practicability of Like-LDA model are validated.

5. **Discussions and Conclusions.** In this paper, we propose an improved topic model by embedding learners' behavioral feature "like" in online course reviews. In this model, the comments learners liked are merged to more accurately analyze the learners' topic interests. This method aims to automatically detect the latent structure of learners' interesting topics in large amounts of course reviews. Through examining the learner-topic and topic-word matrix computed by Like-LDA, their interests can be quickly located and tracked, even more fine-grained items in a visualized way. Experiments are conducted on a real-life data set. Experimental results demonstrate that compared with the benchmark model, Like-LDA can obtain more new hidden topics, higher accuracy of topic detection and words coherency in each topic. These results can effectively help to identify learners'

interests. Employing this proposed model will not only contribute to quickly locating the relevant comments for online learners, but also enable faculties to focus on the meaningful viewpoints and realize personalized recommendation services. In future work, we will fuse operational behaviors, sentiment features and other useful information in the process of text-based mining into the Like-LDA, e.g., replying/forwarding/quoting comments, sentiment symbols, and sentiment orientations of words.

Acknowledgment. This work was supported by the Research funds from the Humanities and Social Sciences Foundation of the Ministry of Education (Grant No. 16YJC880052) and self-determined research funds of CCNU from the colleges' basic research and operation of MOE (Grant Nos. CCNU15A05009, CCNU15A05010, CCNU16JYKX30, CCNU16 A02022). We would like to thank Dr. Wenbin Gan for his constructive comments, who had helped the improvement of this paper.

REFERENCES

[1] S. Hidi and K. A. Renninger, The four phase model of interest development, *Educational Psychologist*, vol.41, no.2, pp.111-127, 2006.

[2] T. M. Nieswandt, Student affect and conceptual understanding in learning chemistry, *Journal of Research in Science Teaching*, vol.44, no.7, pp.908-937, 2007.

[3] J. M. Harackiewicz, K. E. Barron, J. M. Tauer and A. J. Elliot, Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation, *Journal of Educational Psychology*, vol.94, no.3, pp.562-575, 2002.

[4] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2004.

[5] Y. Kim, Y. Park and K. Shim, Digtobi: A recommendation system for digging articles using probabilistic modeling, *WWW 2013*, pp.691-702, 2013.

[6] J. Tang, Z. Meng, X. Nguyen, Q. Mei and M. Zhang, Understanding the limiting factors of topic modeling via posterior contraction analysis, *Proc. of the 31st International Conference on Machine Learning*, pp.190-198, 2014.

[7] I. Titov and R. McDonald, Modeling online reviews with multi-grain topic models, *Proc. of the 17th International Conference on World Wide Web*, pp.111-120, 2008.

[8] S. Moghaddam and M. Ester, ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews, *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.665-674, 2011.

[9] L. Li, L. Zheng, F. Yang et al., Modeling and broadening temporal user interest in personalized news recommendation, *Expert Systems with Applications*, vol.41, no.7, pp.3168-3177, 2014.

[10] B. Jiang and Y. Sha, Modeling temporal dynamics of user interests in online social networks, *Procedia Computer Science*, vol.51, no.1, pp.503-512, 2015.

[11] A. Daud, Using time topic modeling for semantics-based dynamic research interest finding, *Knowledge-Based Systems*, vol.26, pp.154-163, 2012.

[12] A. Ramesh, D. Goldwasser, B. Huang et al., Understanding MOOC discussion forums using seeded LDA, *Proc. of the 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[13] A. Ramesh, S. H. Kumar, J. Foulds et al., Weakly supervised models of aspect-sentiment for online course discussion forums, *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.74-83, 2015.

[14] Z. Yang, J. Guo, K. Cai et al., Understanding retweeting behaviors in social networks, *CIKM*, pp.1633-1636, 2010.

[15] L. Tobarra, A. Robles-Gómez, S. Ros et al., Analyzing the students' behavior and relevant topics in virtual learning communities, *Computers in Human Behavior*, vol.31, pp.659-669, 2014.

[16] M. Qiu, F. Zhu and J. Jiang, It is not just what we say, but how we say them: LDA-based behavior-topic model, *Proc. of the 13th SIAM International Conference on Data Mining*, pp.794-802, 2013.

[17] Z. Zhao, Z. Cheng, L. Hong et al., Improving user topic interest profiles by behavior factorization, *Proc. of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee*, pp.1406-1416, 2015.

[18] D. Ramage, D. Hall, R. Nallapati et al., Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.248-256, 2009.

[19] S. C. Deerwester, S. T. Dumais, T. K. Landauer et al., Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, vol.41, no.6, pp.391-407, 1990.

[20] T. Hofmann, Probabilistic latent semantic indexing, *Proc. of the 22nd Annual International SIGIR Conference*, pp.50-57, 1999.

[21] N. B. Ellison, J. Vitak et al., Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes, *Journal of Computer-Mediated Communication*, vol.19, no.4, pp.855-870, 2014.

[22] L. E. Sherman, A. A. Payton, L. M. Hernandez et al., The power of the like in adolescence effects of peer influence on neural and behavioral responses to social media, *Psychological Science*, pp.1-9, 2016.

[23] M. Steyvers and T. Griffiths, Probabilistic topic models, *Handbook of Latent Semantic Analysis*, vol.7, pp.424-440, 2007.

[24] H. P. Zhang, Q. Liu, X. Q. Cheng et al., Chinese lexical analysis using hierarchical hidden Markov model, *Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing*, vol.17, pp.63-70, 2003.