Participation Prediction and Opinion Formation in MOOC Discussion Forum

Tieying Zhu, Wei Wang, Wei Zhao, and Riming Zhang

***Abstract*—With the development of MOOCs, millions of students enrolled into online courses. The discussion forums in MOOCs provide a virtual community for students to interact with each other. The communication in different topics indicates the engagement of students in the courses and social-learning process during the interactions. In this respect, this paper explores the use of Naive Bayesian classification approach for predicting the participation of the forum and the use of Bayesian-based social-learning approach for modelling the opinion formation process during the discussion and indicating the influence of instructors in the discussion forum. Results on data from 1 Coursera course demonstrate that the poster's retention can be well predicted by Naive Bayes classifier based on the combination of different features of the forum postings; additionally, we find that the superposters may not be the participants who will continue posting in the last several weeks. In terms of social-learning, our analysis indicates participants will aggregate information by repeated interactions and the instructors' post can improve the convergence of learning process to the true belief. These results confirm the influence of the instructors' intervention further.**

***Index Terms*—Discussion forum, MOOCs, participation prediction, opinion formation.**

I. Introduction

Massive open online courses, or MOOCs, have recently become a practical, high-quality and life-long learning platform, which changes the public thinking of education [1]. Multiple providers offer thousands of courses, including the Udacity, Coursera and edX in US, FutureLearn, and Iversity in UK, icourses and xuetangX in China, etc; millions of students enrolled in these courses.

Besides the lecture video, homework and exam, the discussion forum is the indispensable part of MOOCs. As a virtual crowd-sourced learning platform, discussion forum provides a cooperative communication community for students to ask questions about course content and get feedback from peers and instructors. The students' postings in the discussion forum reflect the attention and concern about the course. Recently, the interaction behaviour in the forum, as a data source of learning analysis, is getting more and more attention. These topics not only include correlation between the forum participation and final gains [2]-[5], but also include the instructors' role and intervention in the forum [6]-[8].

In this respect, this paper focuses on the interaction pattern in the discussion forums and tries to answer the two questions: 1) can we predict the learners' retention in forum according to their participation data in earlier weeks? 2) Can we model the opinion aggregation and formulize the influence of instructors in the social learning process in the discussion forum? In order to address these questions, we analyze the social structure of forum and the content of postings, and propose a method to predict the student retention by using Naive Bayesian classification accurately. The results will help to get useful feedback about the learners in time from their early behaviour and improve the understanding of the evolution of discussion forum. It can be used to inspect the relationship with participation retention and performance as well. Further, we use the Bayesian learning theory to analyze the opinion formation and learning process to highlight the influence of instructors in information propagation and aggregation. The findings will provide the direct evidence to show the importance of intervention of the instructors from the point of view of opinion formation in social learning.

II. Related Works

MOOCs are open education platforms, in which the participants are self-motivated to complete courses. In the implementation of MOOCs, one of the concerns is the high drop rate and low student retention. Students' activity data analysis can help to understand learning behaviour, analyze the root causes, and improve the course design. A number of researchers have studied this issue from aspects including course video, assignment, and forum participation, based on the large data sets of MOOCs provided by edX, Coursera, etc [5], [9]-[12].

Besides data of comprehensive course-related activities, some researchers focus on the interaction data behind the discussion forums. Their research usually involved the following aspects: forum evolution, structure and social relation of participants, thread classification, forum participation and performance, and instructors' intervention. One of the earlier works is to convert the forum post-reply data into social interactions to dig the social graph structure and community [13]. For the forum of MOOCs, Yang *et al*. traced the posts and comments of students in 1 course, analyzed students' interactive behaviour and social status in the learning communities using social analysis methods, and predicted the drop using survival model [2]. Rossi & Gnawali

Manuscript received February 20, 2016; revised April 14, 2016. This work was supported in part by the Department of Education of Jilin Province under Grant 2014B04, Department of Science and Technology of Jilin Province under Grant 2015010106JC, 20150204009GX and Technology Foundation for Selected Overseas Chinese Scholar.

Tieying Zhu, Wei Wang, and Wei Zhao are with School of Computer Science, Northeast Normal University, China (e-mail: zhuty@nenu.edu.cn, wangw@nenu.edu.cn, zhaow@nenu.edu.cn).

Riming Zhang is with School of Software Engineering, Northeast Normal University, China (e-mail: zhangrm281@nenu.edu.cn).

examined from the forums of 60 courses to study the relation between the evolution of forum activities and course duration and classify the threads based on language-independent features [14]. Cui & Wise investigated the content-related questions in forums using machine learning method based on linguistic features and found the content-related threads were a minority and under-addressed by instructors [15]. Gillani & Eynon computed the Jaccard index, a similarity indicator, of pair-wise sub-forums participation to reveal the consistency of participation [16]. Their analysis suggests that most forum participation was disparate crowds, not cohesive communities. Huang *et al*. investigate 44 courses in Coursera and found that superposters' activity correlates positively with the forum overall activity [17]. Wang *et al*. used content analysis method to analyze the relationship between students' discussion behaviour and the learning gains [3]. Their results show active and constructive discussion behaviours are significant in predicting students' performance. Klusener *et al*. combined the features of students, like number of answers and up-votes, derived from the forum activities into learning profile in Iversity MOOCs to determine the successful students [4].

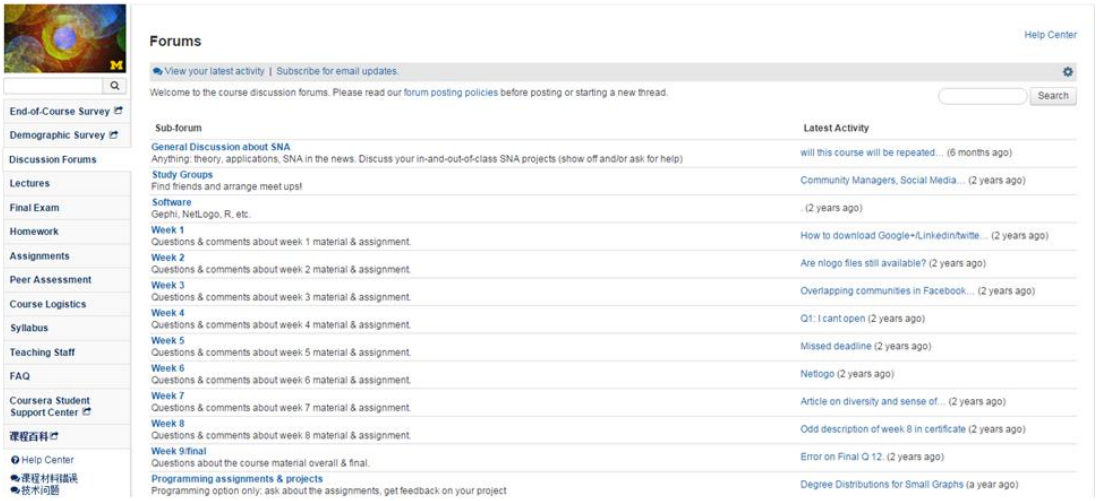

Fig. 1. Information of the discussion forum.

These findings show the positive correlation between the learners' activity and final gains. Compared with these works, this study examines data from the discussion forum with emphasis on forum retention prediction. It can predict who will continue posting in the forum according to the activities in the earlier works and provides a method to get prediction and feedback of learners' personal behaviour in time as well. It is meaningful since it can reflect the evolution of the forum from the personal activities rather than the total number of posts. It also can be combined with the final gains to inspect the relationship with class participation retention and final performance.

For the influence of instructors in discussion forums, Brinton *et al*. analyzed the factors correlated with the discussion volume and participation decline in discussion forums to understand forum activities. Their results show that instructors' active participation can increase the discussion volume but cannot mitigate the participation decline [18]. Chaturvedi *et al*. employed latent categories to abstract contents of posts, and utilized event chain based model to predict the instructors' intervention in MOOC discussion forum [6]. Skrypnyk *et al*. conducted social network analysis on Twitter-based interactions and confirmed that the teachers preserved a high level of influence over the flow of information [7]. Yang *et al*. explored the relationship between confusion and attrition and highlight the need of the intervention [8]. Chandrasekaran *et al*. designed a binary classifier to predict the timing of instructor intervention [19].

In our work, the structure of interaction is extracted as a social network and we adopt the pure Bayesian-learning theory and incorporate with more practical parameters, which are related with the real case, to analysis the convergence of learning process in MOOCs discussion forums, and to highlight the influence of instructors as well.

The related Bayesian social learning research was introduced by Banerjee [20], Bikhchandani *et al*. [21] and Smith and Sorensen [22]. Their model assumed an agent make decisions sequentially based on the observation of his precursor and is known as the sequential social-learning model (SSLM). Gale & Kariv [23], Golub & Jackson [24] and Acemoglu *et al*. [25] extended the sequential structure into social networks and used pure Bayesian-learning to study the uniformity of behaviour. Jadbabaiea *et al*. [26] developed a hybrid model, incorporating a linearly non-Bayesian updating to bridge the gap when the agents fail to adjust their opinion in fully Bayesian manner.

III. Data

In the experiments, we investigate the discussion forum of a 9-week Coursera course provided by Michigan University. We crawled the discussion forum on September 10, 2014. The discussion forum includes 15 sub-forums, according to different topics, such as "General Discussion about SNA", "Study Groups", "software", "Week1", "Week9", etc. The detailed classification information is shown in Fig. 1. Each sub-forum is organized as the temporal sequence of threads according to the time order of the initial post. Each thread has a title and consists of multiple posts and comments. There are 1008 users participated the discussion, and totally publishing 4037 posts, and 2365 comments in 918 threads.

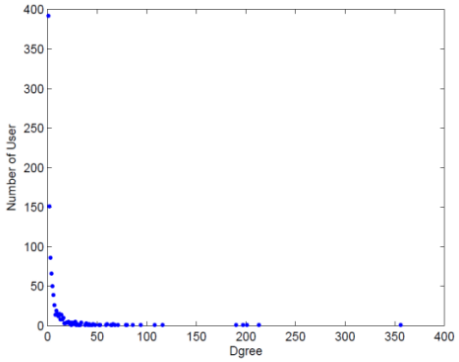

Fig. 2. Node degree distribution.

In this paper, for each post or comment, we treat them equally as a reply to the initial post, since we focus on the message exchanging pattern in the forum. The communication pattern between participants is extracted as a social network. Fig. 2 describes its node-degree distribution. Like in other social networks, it is skewed, known as heavy-tailed or scale-free distribution, in which most nodes have only few links, but, by contrast, there exist some nodes which are extremely linked. In fact, in the discussion forum, 81.8% participants' posting is less than 8, while the highest volume of posting is larger than 350.

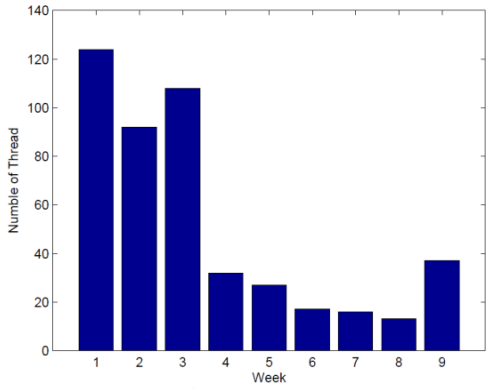

Fig. 3. The number of threads in "Week 1" to "Week 9,"

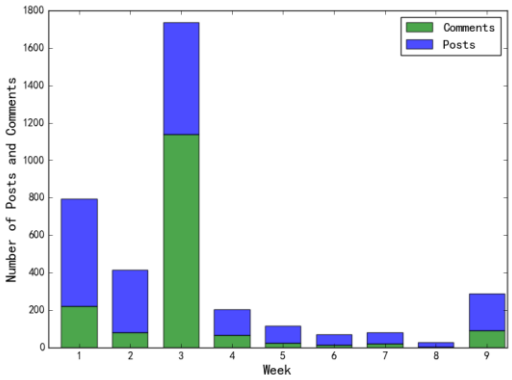

Fig. 4. The number of posts and comments in "Week 1" to "Week 9,"

In the following experiments, we focus on the data accumulated from sub-forum "Week 1" to "Week 9", since there are lots of noisy information in the general and hello sub-forum, and the discussion postings are related with the teaching content during this period. Fig. 3 shows the number of threads in the nine sub-forums. Fig. 4 shows the total number of posts and comments. These two Figures present the declined tendency as a whole in both the number of threads and the volume of postings, although there is a slightly increase in the final week.

IV. Prediction of Participation in Discussion Forum

Bayes' theorem provides a way of calculating the posterior probability from likelihood which is the probability of predictor given class, class prior probability and predictor prior probability. Naive Bayes classifier is a probabilistic classifier based on Bayes' theorem with naive independence assumption between features and it is widely used to deal with classification problems, like email spam. We explore this method to predict whether a user who posts in forums of the first two weeks will continue staying in the forum and posting in the last three weeks. In general context, if the Class C_i represented by a vector $X=(x_1,...,x_n)$ representing n features of an instant, using Bayes' theorem, the conditional probability can be written as $P(C_i \mid X)=\frac{P(X \mid C_i)P(C_i)}{P(X)}$, in which the probability that an instance contains all of the features, given a class C_i, is $P(X \mid C_i)=\prod_{k=1}^{n} P(x_k \mid C_i)$.

Suppose that the participants of the forum are classified into two categories: leaving and continuing, which means those who do not post anything in the last three weeks and those who continue to participate in the discussion. We are going to use the features of a participant' postings in the first two weeks to classify which category he (she) belongs to. Here, four different predictors are used, including whether a posting content contain the high-frequency words, whether a post get a vote, whether the number of his posts is higher than the median of each user, and whether the user is in signature track. Other features may be chosen as predictors too, like the length of posting, frequency of posting. Here we use four features for simplicity. In order to get the high-frequency words, we use Lucene Analyzer [27] to segment words to obtain the high-frequency words in the content of posts. After removing the stop words, the first ten high-frequency words are shown in Table I. Obviously, these words are highly related with the course content.

TABLE I: Highly Frequently Used Words

High-frequency words:
SNA, Gephi, File, Network, Data, Course, Social, nodes, graph, degree

In the experiment, we use 70% of data in the data set as the training set and 30% as testing set. Compared with ground truth of the participation of last three weeks, the prediction accuracy is 86% and the result proves the effectiveness of our method. The method can be used as to predict the participation of each learner and provide an early indication for the evolution of the forum. In fact, there are 673 users participating in the forum "week 1" to "week 9", while there are only 81 users posting in the last three weeks. The result also reflects the low participation retention in the discussion forum.

Further, in order to trace the superposter's and continuing posters' behavior, we illustrate the number of the first three superposters' post and continuing posters' in Fig. 5, and Fig. 6,

respectively. In Fig. 5 and Fig. 6, except the user id 1426602, the instructor assistant, the superposters are not the one who contribute high volume in the forum. Superposters tend to post high volume in the beginning weeks and contributed less in the beginning. The continuing posters did not publish the highest volume of posts, but published relatively constant number of posts in the 9 sub-forums.

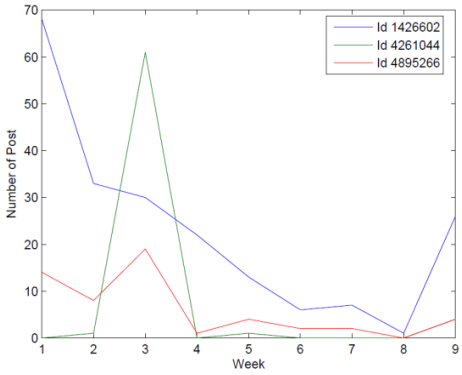

Fig. 5. The number of posts of superposters.

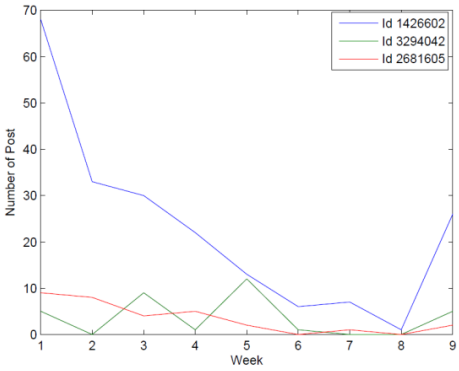

Fig. 6. The number of posts of continuing posters.

V. Social Learning in Discussion Forum

A discussion forum is a type of asynchronous learning network used to increase out-of-class student dialogue about course content. It also can be treated as the social learning process in the corporative environment, in which the initiator puts forward a question and the followers publish their view based on their own belief and their observation of the previous posting, until they get the uniform opinion. Can we formulize the aggregation process of different opinions and indicate influence of instructors in the social learning process? In the asynchronous discussions, each participant can repeatedly post his own opinions and can get the full knowledge of the precursors.

In order to answer this question, we use Bayesian learning model to characterize the convergence process in a socially connected world. Bayesian social learning model focuses on formulating a dynamic model of opinion formation in social networks in which users repeatedly update their personal belief incorporating the views of their neighbours using Bayesian updating rule, i.e. the personal belief is adjusted by its own personal signal, network structure and the observation of the other behaviours.

A. Network Model of the Discussion Forum

We describe the discussion forum using a directed graph $G=(V,E)$, where V is the vertex set and E is the edge set. Each vertex corresponds to a participants, and an edge $e_{i,j}$ captures the fact that participant i follow a message the post of participant j. For each participant i, define $N_i=\{j\in V, e_{i,j}\in E\}$ is the neighbor set of i. The adjacent matrix of graph G, $A=(\alpha_{ij})_{n\times n}, a_{i,j}>0$, is the weights that the participant assigns to the Bayesian posterior conditional on the private or neighbor's signal, where $a_{i,j}$ is the influence or reliance of node j to its neighbor node i, when $i\neq j$, otherwise $a_{i,i}$ is the self-reliance. The weights of $a_{i,j}$ must satisfy $\sum_{j\in N_i\cup\{i\}} a_{i,j}=1$. In this paper, we define the reliance weight is related with the interactions between nodes. More precisely, we define $a_{i,j}$ as:

$$a_{i,j}=\lambda+1/2*C_j/C, \tag{1}$$

where C_j is the interaction frequency with node j, and C is the total interaction times with all the nodes. λ is a threshold for two types of participants: students and instructors. Instructors should have higher λ value than students because of the higher influence of instructors. For example, $\lambda=0.3$ for students, while $\lambda=0.5$ for instructors..

B. Belief and Outside signals

Let $\Theta=\{\theta_1,\theta_2,...,\theta_n\}$ denote a finite state space of the social event in the discussion forum, and the underlying true state $\theta^*\in\Theta$. Participant i's belief about state θ_k at time t is the probability $\mu_{i,t}(\theta_k)$, and $0<\mu_{i,t}(\theta_k)\leq 1$ with the constraint of $\sum_k \mu_{i,t}(\theta_k)=1$. In the case of discussion forum, we have two states: θ_1 or θ_2 to represent the true or false, respectively, i.e., $\Theta=\{\theta_1,\theta_2\}$, in which the true state $\theta^*=\theta_1$. Initial beliefs of any participant i is defined as function of the total number of his postings n, and written as $\mu_{i,0}(\theta_1)=0.5^{1/\lg n}$. And $\mu_{i,0}(\theta_2)=1-\mu_{i,0}(\theta_1)$.

For the action, a participant can generally improve his decision by observing what others have done as an outside signal before choosing his own action. It means agents can observe one another's actions and it is rational for them to learn from one other (Gale and Kariv, 2003). Suppose that the set of actions is $A=\{A_1,A_2\}$, in which the correct action $A^*=A_1$. The action of participant i about state θ at time t is the likelihood function $l_i(A_{m,t}\mid\theta)$, satisfying $\sum_m l(A_m\mid\theta)=1$, i.e. $l(A_2\mid\theta_1)=1-l(A_1\mid\theta_1)$. For any participant, we define the likelihood of doing a correct action conditional on a true state as a function of the ratio between

voted posts and total posts:

$$l(A_1 \mid \theta_1) = 0.5 + 0.5 * \frac{p_v}{p_s}, \quad (2)$$

where p_v is the number of his postings being voted, and p_s is the all the postings of this participant. Further, We assume the likelihood of doing a wrong action conditional on a false state is equal to the likelihood of doing a correct action conditional on a correct state, i.e., $l(A_2 \mid \theta_2) = l(A_1 \mid \theta_1)$.

C. Belief Update and Probability of Different Actions

We use the Bayesian formula for the Bayesian belief updates. For a single participant, Bayesian posterior belief based on the outside signal s observing, can be written as follows [26]: $\mu(\theta \mid s) = \frac{l(s \mid \theta)\mu(\theta)}{m(s)}$, where $m(s) = \sum_{\theta \in \Theta} l(s \mid \theta)\mu(\theta)$.

For multiple participants, we define the posterior belief $\mu_{i,t+1}(\theta)$ as follows:

$$\mu_{i,t+1}(\theta) \square \mu_{i,t+1}(\theta \mid A_{j,t}) = \frac{\sum_{j \in N} a_{i,j} l(A_{j,t} \mid \theta)\mu_{j,t}(\theta)}{\sum_{\theta \in \Theta} l(A_{j,t} \mid \theta)\mu_{j,t}(\theta)} \quad (3)$$

where $A_{j,t}$ is the action taken by participant j at time t.

Each participant will take an action according to its belief to the discussion question according to its belief μ at time t. Compared with the common participant, the instructor or teaching assistant can make correct judgment and do right action with higher probability. For each common participant, the probability of doing different actions is given by: $P(A_{i,t}) = \sum_{\Theta} l(A_{i,t} \mid \theta)\mu_{i,t}(\theta)$. For two types of actions,

$$P(A_1) = \mu(\theta_1) l(A_1 \mid \theta_1) + \mu(\theta_2) l(A_1 \mid \theta_2),$$

$$P(A_2) = \mu(\theta_1) l(A_2 \mid \theta_1) + \mu(\theta_2) l(A_2 \mid \theta_2) \quad (4)$$

For each instructors or teaching assistants, the probability of doing actions is adopted randomly in interval $[0..1]$, satisfying

$$l(A_1 \mid \theta_1) < P(A_1) < 1 \text{ and } 0 < P(A_2) < l(A_1 \mid \theta_1) \quad (5)$$

D. Convergence of Bayesian Learning

We will now analyze how the opinions aggregate in the cooperative learning environment. We explore the model and parameters shown in last Section to one of the threads in the course forum, which has six participants, user1, user2, user3, user4, user5 and user 6. User 4 is the teaching assistant. According to the data set we use, we first get the following information of the six participants: 1) totally number of postings, which is shown in vector V, 2) the voted postings, which is shown in vector W, 3) interaction times between each pair, which is shown in Matrix T. Then, the initial belief of each user $\mu_{i,0}(\theta_1)$ and the reliance weight between the six users $a_{i,j}$ can be computed by Eq. (1), based on 1) and 2). We use the expressions shown in Eq. (2) to compute the likelihood of participant i's action about state θ at time t based on 2) and 3).

$$T = \begin{pmatrix} 1 & 0 & 0 & 2 & 0 & 1 \\ 0 & 3 & 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 5 & 1 & 1 \\ 0 & 2 & 0 & 1 & 13 & 1 \\ 1 & 1 & 1 & 1 & 1 & 7 \end{pmatrix},$$

$V = [4, 58, 3, 333, 34, 11]$, $W = [0, 37, 2, 103, 9, 1]$.

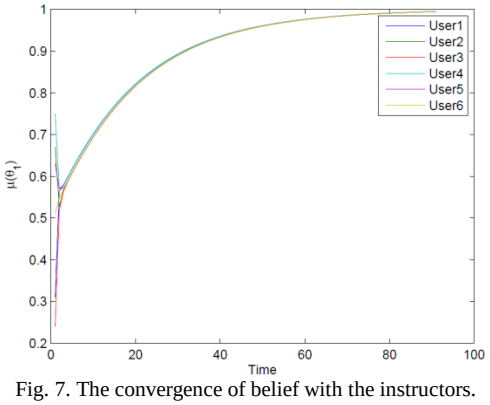

Fig. 7. The convergence of belief with the instructors.

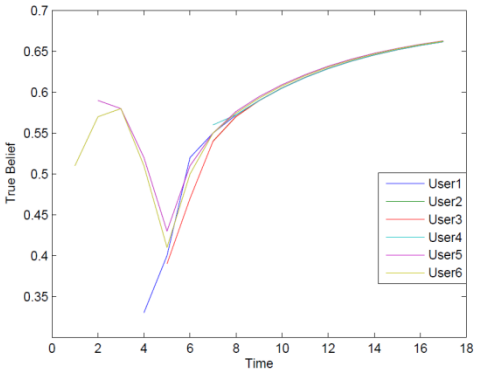

Fig. 8. The detailed convergence of belief with instructors.

Given a new post is published at time t, each participant's belief is updated according to Eq. (3), and the probability of taking different actions is computed as Eq. (4) and (5). Here, we use the expectation of belief to evaluate the evolution of belief for each state, since we do not analysis the content of each post to see whether it is a right answer or not. Fig. 7 illustrates the belief about true state θ_1 moves towards 1, as the observed information increases. In order to see the detailed information of the convergence, Fig. 8 shows the beginning of the opinion aggregation. With the intervention of

User 4 at Time 7, the belief is gradually lager than 0.5 and close to 1. In order to see the influence of the instructors, we replace participant 4 with a common students by changing its probability of giving right answer to the common students, and keep the other interactions' and posts' parameters. The results are shown in Fig. 9 and Fig. 10. From these two figures, we can see that the convergence still exists, but the expectation of true belief is between 0.5 and 0.6. The results show the instructor can helps mitigate the confusions in the discussion and confirm the influence of instructions' in the discussion forums.

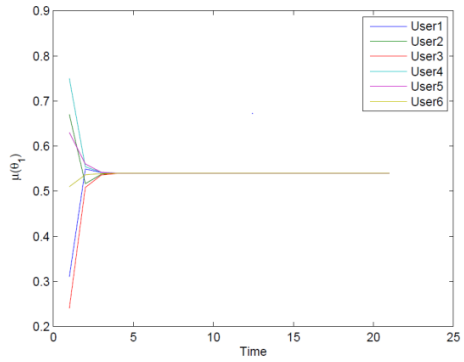

Fig. 9. The convergence of belief without the instructors.

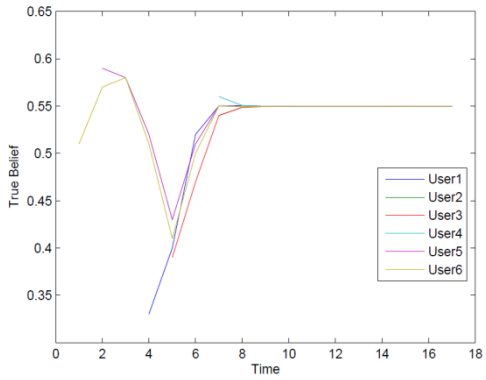

Fig. 10. The detailed convergence of belief without instructors.

VI. Conclusion

In this paper, we focus on the participation retention and social learning process in discussion forum of MOOCs. First, we explore the application of Naive Bayesian classifier in the prediction of continuing posters in the last weeks, based on features in the participant's postings in first two weeks. The experiment results show the accuracy of prediction. Additionally, we find that the superposters may not be the continuing posters, although they are the higher-volume contributors in the forum. These findings will be helpful for understanding each forum participant activities, including the superposters, and evolution of discussion forum from the view of personal behaviour and get useful information to increase the forum engagement.

Further, we exploit the social-learning process by adopting Bayesian learning model. Our results show that the process is convergent whether there is the intervention of instructors or not. It means participants can aggregate opinion over the discussion forum. But with the discussion of instructors, the expectation of true belief about a decision is close to 1. These results emphasize the importance of instructions' opinion. This research is an early work to study the social-learning process in discussion forums. For future work, we are interested in the content analysis of the postings to track the social-learning process more accurately.

References

[1] M. Waldrop, "Online learning: Campus 2.0," *Nature*, vol. 495, no. 7440, pp. 160-163, 2013.
[2] D. Yang, T. Sinha, D. Adamson, and C. P. Rose, "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses," presented at NIPS Workshop on Data Driven Education, Dec, 2013.
[3] X. Wang *et al*., "Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains," presented at 8th International Conference on Educational Data Mining (EDM), Madrid, Spain, June, 2015.
[4] M. Klusener and A. Fortenbacher, "Predicting students' success based on forum activities in MOOCs," in *Proc. IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pp. 925–928, 2015.
[5] Z. Jiang, Y. Zhang, and X. Li, "Learning behavior analysis and prediction based on MOOC data," *Journal of Computer Research and Development*, 2015, vol. 52, no. 3, pp. 614-628.
[6] S. Chaturvedi, D. Goldwasser, and H. Daumé III, "Predicting instructor's intervention in MOOC forums," *ACL*, vol. 1, 2014, pp. 1501-1511.
[7] O. Skrypnyk, S. Joksimović, V. Kovanović *et al*., "Roles of course facilitators, learners, and technology in the flow of information of a cMOOC," *International Review of Research in Open and Distributed Learning*, vol. 16, no. 3, 2015, pp. 188-217.
[8] D. Yang *et al*., "Exploring the effect of confusion in discussion forums of massive open online courses," presented at the Second ACM Conference on Learning @ Scale Conference, New York, NY, USA, 2015.
[9] A. Ramesh *et al*., "Modeling learner engagement in MOOCs using probabilistic soft logic," presented at NIPS Workshop on Data Driven Education, 2013.
[10] A. Anderson *et al*., "Engaging with massive online courses," in *Proc. the 23rd International Conference on World Wide Web (WWW'14)*, New York, NY, pp. 687-698.
[11] D. Seaton *et al*., "Who does what in a massive open online course?" *Communications of the ACM*, vol. 57, no. 4, pp. 58–65, 2014.
[12] S. Jiang *et al*., "Predicting MOOC performance with week 1 behavior," in *Proc. the 7th International Conference on Educational Data Mining (EDM)*, pp. 273-275, 2014.
[13] S. Dawson, A. Bakharia, and E. Heathcote, "SNAPP: Realising the affordances of real-time SNA within networked learning environments," in *Proc. the 1st International Conference on Learning Analytics and Knowledge (LAK '11)*, ACM, New York, NY, pp. 168-173, 2011.
[14] L. A. Rossi and O. Gnawali, "Language independent analysis and classification of discussion threads in Coursera MOOC forums," presented at the IEEE International Conference on Information Reuse and Integration, 2014.
[15] Y. Cui and A. F. Wise, "Identifying content-related threads in MOOC discussion forums," in *Proc. Learning @ Scale*, pp. 209–303, 2015.
[16] N. Gillani and R. Eynon, "Communication patterns in massively open online courses," *Internet and Higher Education*, vol. 23, pp. 18-26, 2014.
[17] J. Huang *et al*., "Superposter behavior in MOOC forums," in *Proc. the First ACM conference on Learning @ Scale Conference*, New York, USA, pp. 117-126, 2014.
[18] C. G. Brinton *et al*., "Learning about social learning in MOOCs: From statistical analysis to generative model," 2013.
[19] M. K. Chandrasekaran *et al*., "Learning instructor intervention from MOOC forums: Early results and issues," in *Proc. the 8th International Conference on Education Data Mining (EDM)*, Madrid, Spain, pp. 218-225, 2015.
[20] A. Banerjee, "A simple model of herd behavior," *Quarterly Journal of Economics*, vol. 107, no. 3, pp. 797-817, 1992.
[21] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascade," *Journal of Political Economy*, vol. 100, no. 5, pp. 992-1026, 1992.
[22] L. Smith and P. Sorensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, pp. 371–398, 2000.

[23] D. Gale and S. Kariv, "Bayesian learning in social networks," *Games and Economic Behavior*, vol. 45, pp. 329–346, 2003.
[24] B. Golub and M. O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, vol. 2, no. 1, pp. 112-149, 2010.
[25] D. Acemoglu *et al*., "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201-1236, 2011.
[26] A. Jadbabaiea *et al*., "Non-bayesian social learning," *Games and Economic Behavior*, vol. 76, issue 1, September 2012, pp. 210–225.
[27] Lucene analyzer. [Online]. Available: http://lucene.apache.org/core/4_3_0/core/org/apache/lucene/analysis/Analyzer.html

Tieying Zhu received the B.S. degree in information management from Northeast Normal University, Changchun, China, in 1996, and the M.S. and Ph.D. degrees in computer science from Jilin University, Changchun, in 1999 and 2005, respectively. Since 1999, she has been with the School of Computer Science, Northeast Normal University, where she is currently an associate professor. Her research interests include complex networks, opportunistic mobile social networks, and network security.

Wei Wang received the B.S. degree in computer science and technology, from Northeast Normal University, Changchun, China, 2014. He is a postgraduate student of School of Computer Science, Northeast Normal University. His research interest includes social network analysis and learning analyst.

Wei Zhao received the B.S. and M. S. degree in physics from Northeast Normal University, Changchun, China, in 1986 and 1989, respectively. Since 1989, she has been with the School of Computer Science, Northeast Normal University, where she is currently a professor and Ph. D supervisor. Her research interests include distance education and cooperative learning and learning analyst.

Riming Zhang received the B. S. and M. S. degree in automation from Jilin University of Science and Technology, Changchun, China in 1994 and1997, the Ph.D from Jilin University, Changchun, in 2008. Since 2005, she has been with the School of Software Engineering, Northeast Normal University, where she is currently an associate professor. Her research interest includes social network analysis, data mining and learning analyst.