Influence Analysis by Heterogeneous Network in MOOC Forums: What can We Discover?

Zhuoxuan Jiang[1], Yan Zhang[2], Chi Liu[1], Xiaoming Li[1]
Institute of Network Computing and Information System
Peking University, Beijing, China
[1]{jzhx,liuchi,lxm}@pku.edu.cn, [2]zhy@cis.pku.edu.cn

ABSTRACT

With the development of Massive Open Online Courses (MOOC) in recent years, discussion forums there have become one of the most important components for both students and instructors to widely exchange ideas. And actually MOOC forums play the role of social learning media for knowledge propagation. In order to further understand the emerging learning settings, we explore the social relationship there by modeling the forum as a heterogeneous network with theories of social network analysis. We discover a specific group of students, named representative students, who feature large engagement in discussions and large aggregation of the majority of the whole forum participation, except the large learning behavior or the best performance. Based on these discoveries, to answer representative students' threads preferentially could not only save time for instructors to choose target posts from all, but also could propagate the knowledge as widespread as possible. Furthermore if extra attention is paid to representative students in the sight of their behavior, performance and posts, instructors could readily get feedback of the teaching quality, realize the major concerns in forums, and then make measures to improve the teaching program. We also develop a real-time and effective visualization tool to help instructors achieve these.

Keywords

MOOC forum, Coursera, influence, behavior, performance, heterogeneous network

1. INTRODUCTION

Comparing with the traditional distance education or online courses, discussion forums in Massive Open Online Courses (MOOC) offer a big and lively venue for communication between students and instructors, which has been proved important for large-scale social learning [1, 7, 9]. However, due to their massiveness, the forums are full of various information relevant and irrelevant to the course [6]. So how to fast and accurately extract valuable information from the large-scale settings has become a problem to which priority should be given.

Considering Twitter, Facebook or StackOverflow, MOOC forums look similar to some kind of social media because of the large number of participants and their interactivity. Every member in the forum may talk about course content, such as asking or answering a question. The intensive interaction between them actually supports the knowledge propagation between members of the learning community. However here comes up a dilemma. In light of knowledge propagation, the proportion of instructors' responses is expected as large as possible in order to resolve students' questions; But considering the scale, instructors could not have enough time to read every thread. In order to cope with this situation, we propose a trade-off solution that extracts influential students from all and recommended them to instructors. Then instructors could make decisions in a much smaller scale and their' effort would be amplified based on principles of influence propagation [12, 16, 24].

Although the definition of influence is various from different perspectives, we leave aside others except instructor for the time being in this paper. We conceive in each forum there could be a group of influential students who attract many others to interact with them, just like the verified accounts in Twitter. We call them 'representative students' and they involuntarily undertake the responsibility for knowledge propagation. So instructors could amplify the influence of right answers by preferentially responding to questions of representative students. Thus, many more students who pay attention to representative students' answers would also benefit without actually having a response by the instructor. On the other hand, given that representative students' threads may get a lot of attention, instructors could address the main concerns in the learning community more promptly. Through the rank list of representative students' influence, the chief instructor could also realize whether other instructors (or called TAs) are on duty, since TAs' influence could be calculated meanwhile. As we show later in this paper, representative students' performance is not the best within the learning community, but given their positive motivation and high volume of messages answering promptly their questions is beneficial for the whole learning community.

Since posts irrelevant to the course are unavoidable in such a free forum, for example chatting, making friends or other things, it is not reasonable to directly regard superposter [9] as representative students or merely consider their social relationship. Experiments later in this paper approve the opinion and find post contents are useful. That being the case, since we regard the interaction in MOOC forums as the procedure of knowledge propagation in social media, we could build a heterogeneous network [23] to model the forum with two kinds of entities by leveraging theories of networked entities ranking. Then we can get a rank list of students' in-

fluence from that network with a specially designed algorithm. The higher a student ranks on the list, the more influential she would be. This model could fully utilize the social information and textual messages to avoid outliers or exceptions (e.g. someone who always submits posts irrelevant to the course).

To our knowledge, this is the first work to adopt a heterogeneous network to model social relationship in MOOC forums and extract representative students. We also propose a novel algorithm for ranking students' influence based on graphic theories. Experimental results show the effectiveness and efficiency of the algorithm are both decent. Through the analysis of representative students' log data, we find they engage highly and aggregate much participation except the excellent grades, which suggests they are representative for instructors to watch the class and are the first low hanging fruit for increasing the passing rate. Analysis of historical records of interaction between instructors and students indicates it is time-saving and meaningful for instructors to recommend threads of representative students. Based on those discoveries, we developed a web service of visualization tool as an assistant for instructors to achieve the conception of supervising their class effort-savingly.

2. RELATED WORK

In traditional off-line classes, the scale is relatively small and face-to-face Q&A is not a challenge. And in traditional online education or online video class, not only the scale is not large enough but the absence of instructors is very common. However, a widespread viewpoint is that it is quite important for MOOC to make students engage in a social learning environment to guarantee and improve the teaching quality [1, 6, 7, 18].

In view of researches in the field of Community Question Answering (CQA), issues related to this paper are about expert finding and forum search [21]. Recently, several novel methods for finding experts in CQA have been provided [26, 29, 30]. Nevertheless, there would be rare experts in MOOC forum due to the specificity that a MOOC forum is not open to all kinds of discussions and it just belongs to the corresponding course for students to acquire knowledge. Also the definition of representative students here is different from that of experts. On the other hand, the task of discovering representative learners and their posts seems like forum search [3, 19] which develops a mechanism analogous to a search engine. But here we concentrate on just the ranking result and not emphasise the accuracy of retrieval. Except those general forum-related work, recently some researches of MOOC forums have been published from various perspectives. For example, Yang et al. [25] tried thread recommendation for MOOC students with method of an adaptive feature-based matrix factorization framework. Wen et al. [22] analyzed the sentiment in MOOC forums via students' words for monitoring their trending opinions. And Stump et al. [20] proposed a framework to classify forum posts.

The classical PageRank [5] and HITS [14] have been applied on broad problems of networked entities ranking and been promoted to solve problems in heterogeneous network [11, 15, 27]. [17, 28] built a heterogeneous network with two types of nodes to discover the influential authors with scientific repository data, which is similar to our work. The point in common is to discover influential entities with iteration by building a graphic model. In this paper, we leverage that principle and build a new heterogeneous network to model MOOC forum and discover representative students.

Besides, many MOOC log analysis also involve forums. Anderson et al. [2] deployed a system of badges to produce incentives for activity and contribution in the forum based on behavior patterns. Huang et al. [9] specially analyzed the behavior of superposter in 44 MOOC forums and found MOOC forums are mostly healthy. Kizilcec et al. [13] did a research on the behavior of students disengagement. Some technical reports and study case papers also involved behavior analysis of MOOC students in forums, such as [8] and [4]. Nevertheless, we believe incentives established on intelligent analysis of various data like social information and textual messages would be more reasonable than on the pure credits mechanism in traditional forums, since the latter only considers the quantity of behavior while not the quality.

Table 1: Pairs of course code and course title

Course Code	Course Title
peopleandnetworks-001	Networks and Crowds
arthistory-001	Art History
dsalgo-001	Data Structures and Algorithms A
pkuic-001	Introduction to Computing
aoo-001	The Advanced Object-Oriented Technology
bdsalgo-001	Data Structures and Algorithms B
criminallaw-001	Criminal Law
pkupop-001	Practice on Programming
chemistry-001	General Chemistry (Session 1)
chemistry-002	General Chemistry (Session 2)
pkubioinfo-001	Bioinformatics: Introduction and Methods (Session 1)
pkubioinfo-002	Bioinformatics: Introduction and Methods (Session 2)

Table 2: Statistics per course

Course	# threads	# posts	# votes
peopleandnetworks-001	219	1,206	304
arthistory-001	273	2,181	1,541
dsalgo-001	283	1,221	266
pkuic-001	1,029	5,942	595
aoo-001	97	515	204
bdsalgo-001	319	1,299	132
criminallaw-001	118	763	648
pkupop-001	1,085	6,443	977
chemistry-001	110	591	65
chemistry-002	167	715	678
pkubioinfo-001	361	2,139	1,474
pkubioinfo-002	170	942	235
Overall	4,259	24,042	-

3. DATASET

We use all the log data of 12 courses from Coursera platform. They were offered in Fall Semester of 2013 and Spring Semester of 2014. There are totally over 4,000 threads and over 24,000 posts. For convenience later in the paper, Table 1 lists the pairs of course code and course title. Table 2 shows the statistics of the dataset per course. Here posts denotes responses including posts and comments. We can see both the subjects and scales range widely.

4. MODEL AND ALGORITHM

In order to model MOOC forums as social media, the first challenge is that no explicit post-reply relationship which describes who replies who is recorded. We simplify this problem and assume

Table 3: Attributes of the heterogeneous network constructed per course

Course	G_S			G_K			G_{SK}	
	n_S	$\|E_S\|$	$\|E_S\|/n_S^2$	n_K	$\|E_K\|$	$\|E_K\|/n_K^2$	$\|E_{SK}\|$	$\|E_{SK}\|/(n_S+n_K)^2$
peopleandnetworks-001	321	3,287	0.032	1,193	104,821	0.074	4,814	0.002
arthistory-001	540	17,022	0.058	3,376	1,019,289	0.089	14,195	0.001
dsalgo-001	295	1,876	0.022	1,152	124,118	0.094	5,009	0.002
pkuic-001	768	19,801	0.034	2,302	302,989	0.057	14,599	0.002
aoo-001	175	1,963	0.064	783	73,208	0.119	2,597	0.003
bdsalgo-001	225	2,369	0.047	781	23,540	0.039	3,133	0.003
criminallaw-001	219	2,971	0.062	1,224	123,737	0.083	4,577	0.002
pkupop-001	628	12,883	0.033	1,748	88,035	0.029	13,807	0.002
chemistry-001	130	886	0.052	1,055	111,026	0.100	2,685	0.002
chemistry-002	125	2,341	0.150	964	61,425	0.066	2,574	0.002
pkubioinfo-001	594	22,275	0.063	686	46,768	0.099	1,946	0.001
pkubioinfo-002	189	1746	0.049	380	16662	0.115	784	0.002

Table 4: Notations

Notation	Description
$G = (V, E, W)$	heterogenous network
$G_S = (V_S, E_S, W_S)$	student subnetwork
$G_K = (V_K, E_K, W_K)$	keyword subnetwork
$G_{SK} = (V_{SK}, E_{SK}, W_{SK})$	bipartite subnetwork
n_S, n_K	$\|V_S\|$, $\|V_K\|$

if two students appear in the same thread, they have the same topic interests and the one whose post is chronologically later replies the other. As mentioned in previous sections, post contents of representative students should be course-related. Thus it may be not enough to cover that demand with only extracting the post-reply relationship. Based on the fact that the most post contents are course-related [9], we add the keywords as another kind of entities into the model to construct the heterogenous network. The keywords here are all meaningful nouns in post contents and they could represent various aspects of topics. Other kinds of parts of speech are unexplored at the present. The role of keywords in the heterogenous network is to help the algorithm reinforce the influence of students who involve more topics, which ensures the need that posts of presentative students are course-related. Figure 1 shows the demo of the heterogeneous network, and Table 4 lists the defined notations.

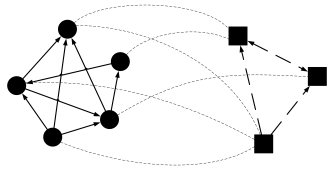

Figure 1: Demo of the heterogeneous network G. Circles denote V_S and rectangles denote V_K. Solid lines with arrows denote the co-presence relationship between students in the same thread and arrows denote one whose post is later points to the other.Dash lines with arrows denote the co-presence of keywords in the same thread but directed or bidirectional arrows mean the two keywords are in the different post or not. Dash lines without arrows denote the authorship between students and keywords. The weight values mean the times of co-presence of two entities on corresponding edges. Self co-presence is meaningless and all ignored.

This model captures the characteristic that representative students would own more latent post-reply relationship and involve more topics. After building the network through log dataset, the basic attributes of graphs per course are calculated (Table 3).

For co-ranking students and keywords, we need an algorithm. We simulates two random surfers jumping and walking in the heterogeneous network and design the algorithm named Jump-Random-Walk (JRW). We assume the weights W represent the influence between entities and the algorithm's task is to discover the most influential students, namely representative students. Figure 2 shows the framework of JRW algorithm.

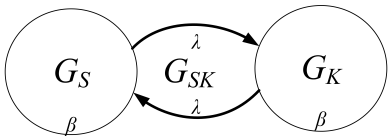

Figure 2: The framework of Jump-Random-Walk algorithm. β is the probability of walking along an edge within G_S or G_K. λ is the probability for jumping from G_S to G_K or in reverse. $\lambda = 0$ means to discover representative students only by using post-reply relationship. We assume the probabilities of each jump or walk are consistent.

Denote $\mathbf{s} \in \mathbb{R}^{n_S}$ and $\mathbf{k} \in \mathbb{R}^{n_K}$ are the ranking result vectors, also probability distributions, whose entries are corresponding to entities of V_S and V_K, subject to $\|\mathbf{s}\|_1 \leq 1$ and $\|\mathbf{k}\|_1 \leq 1$. Denote the four transition matrixes, G_S, G_K, G_{SK} and G_{KS}, for iteration as $S \in \mathbb{R}^{n_S \times n_S}$, $K \in \mathbb{R}^{n_K \times n_K}$, $SK \in \mathbb{R}^{n_{SK} \times n_{SK}}$, and $KS \in \mathbb{R}^{n_K \times n_S}$ respectively. Adding the probability of random jumping for avoiding trapped in connected subgraph or set of no-out-degree entities, the iteration functions are

$$\mathbf{s} = (1-\lambda)(\beta S\tilde{\mathbf{s}} + (1-\beta)\mathbf{e}_{n_S}/n_S) + \lambda SK\tilde{\mathbf{k}}, \quad (1)$$

$$\mathbf{k} = (1-\lambda)(\beta K\tilde{\mathbf{k}} + (1-\beta)\mathbf{e}_{n_K}/n_K) + \lambda KS\tilde{\mathbf{s}}, \quad (2)$$

where $\mathbf{e}_{n_S} \in \mathbb{R}^{n_S}$ and $\mathbf{e}_{n_K} \in \mathbb{R}^{n_K}$ are the vectors whose all entries are 1. The mathematical forms of four transition matrixes are

$$S_{i,j} = \frac{w^S_{i,j}}{\sum_i w^S_{i,j}} \quad where \sum_i w^S_{i,j} \neq 0, \quad (3)$$

$$K_{i,j} = \frac{w^K_{i,j}}{\sum_i w^K_{i,j}} \quad where \sum_i w^K_{i,j} \neq 0, \quad (4)$$

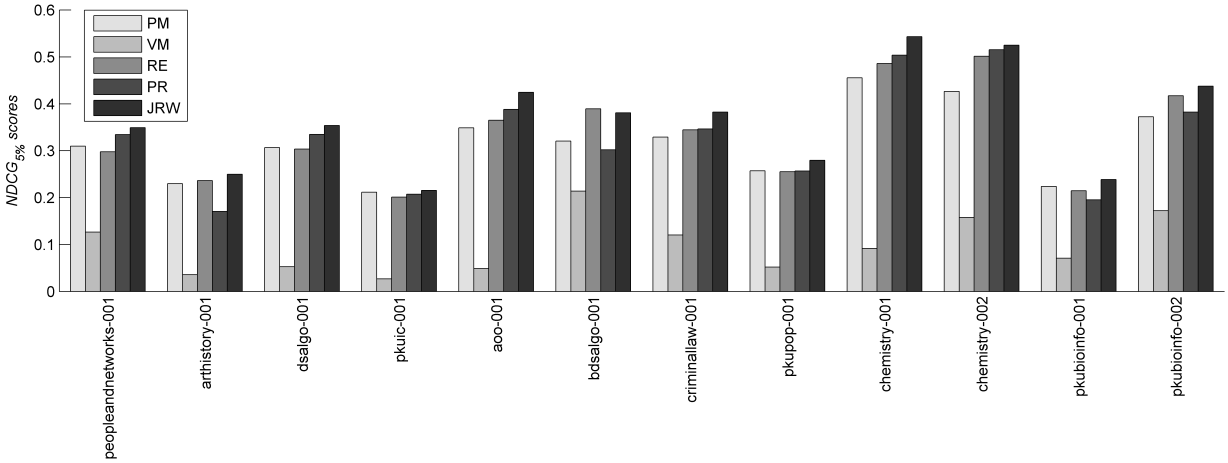

Figure 3: $NDCG_{5\%}$ scores of different rankings

$$SK_{i,j} = \frac{w_{i,j}^{SK}}{\sum_i w_{i,j}^{SK}}, \tag{5}$$

$$KS_{i,j} = \frac{w_{i,j}^{KS}}{\sum_i w_{i,j}^{KS}} \quad where \sum_i w_{i,j}^{KS} \neq 0. \tag{6}$$

$w_{i,j}^S$ is the weight of the edge from V_i^S to V_j^S, $w_{i,j}^K$ is the weight of the edge between V_i^K and V_j^K, $w_{i,j}^{SK}$ is the weight of the edge between V_i^S and V_j^K and $w_{i,j}^{KS}$ is the weight of the edge between V_i^K and V_j^S. Actually $w_{i,j}^{SK} = w_{j,i}^{KS}$. When $\sum_i w_{i,j}^S = 0$, it means the student V_j^S is always the last one in a thread. If $\sum_i w_{i,j}^K = 0$, it means the keyword V_j^K always has no peer in a thread. Actually this situation almost never happens in our filtered data. $\sum_i w_{i,j}^{SK} = 0$ is also impossible, which means every keyword would have at least one author (student). On the contrary, it does not make sure that every student would post at least one keyword, because maybe there is some post having nothing valuable or not containing any nounal keyword. Algorithm 1 shows the detail of JRW algorithm below.

Algorithm 1 Jump-Random-Walk on G

INPUT $S, K, SK, KS, \beta, \lambda, \epsilon$
$\mathbf{s} \leftarrow \mathbf{e}/n_S$
$\mathbf{k} \leftarrow \mathbf{e}/n_K$
repeat
  $\tilde{\mathbf{s}} \leftarrow \mathbf{s}$
  $\tilde{\mathbf{k}} \leftarrow \mathbf{k}$
  $\mathbf{s} = (1-\lambda)(\beta S\tilde{\mathbf{s}} + (1-\beta)\mathbf{e}_{n_S}/n_S) + \lambda SK\tilde{\mathbf{k}}$
  $\mathbf{k} = (1-\lambda)(\beta K\tilde{\mathbf{k}} + (1-\beta)\mathbf{e}_{n_K}/n_K) + \lambda KS\tilde{\mathbf{s}}$
until $|\mathbf{s} - \tilde{\mathbf{s}}| \leq \epsilon$
return $\mathbf{s}, \mathbf{k}$

5. EXPERIMENTS

We do not exclude the data of instructors (or TAs) and regard everyone in the forums as 'students'. So that instructors' influence can also be evaluated in the uniform framework. Since the courses are all in Chinese and the contents are overwhelmingly most in simple

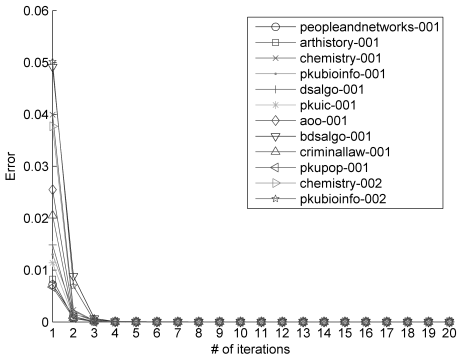

Figure 4: Iteration speed of Jump-Random-Walk

Chinese or traditional Chinese, we filter the non-Chinese contents in the preprocessing step with a tool of Chinese words segmentation which is essential for extracting Chinese keywords. Also we filter the HTML tags irregularly existed. During this process, most spam and valueless posts are filtered incidentally.

To evaluate the effectiveness of JRW, we set some competitors listed below.

- **Post the most** (PM), for superposter by quantity. The more amount and frequency of posts are submitted, the higher she would rank.
- **Be voted the most** (VM), for superposter by quality. The larger ratio of the number of votes earned to the average number of votes in a forum, the higher she would rank.
- **Reputation** (RE), for superposter by reputation. It is a reputation score maintained by the Coursera platform and can be seen as a measure of both the quantity and quality of a forum student's contribution.
- **PageRank** (PR), for representative students only by post-reply relationship. It computes each forum student's influence only in G_S with PageRank algorithm.

Table 5: Representative students' behavior and performance. $P(R|T)$ is the proportion of the number of threads initiated by representative students to the all. $P(R|P)$ is the proportion of the number of posts by representative students to the all. *Over Rate* is the deviation of the average numbers of posts per thread initiated by representative students and the all. $P(R|V)$ is the proportion of the number of watching video by representative students to the all. $P(R|Q)$ is the proportion of the number of submitting quiz by representative students to the all. $P(R|C)$ and $P(R|C,D)$ are the proportions of certificated representative students and certificated representative students with distinction to the all. *Precise* is the proportion of the number of posts by instructors in threads initiated by representative students to that of all the instructors' posts. *Recall* is the proportion of the number of threads replied by instructors to that of threads initiated by representative students.

Course	Forum Behavior			Learning Behavior		Performance		Instructor	
	$P(R\|T)$	$P(R\|P)$	*Over Rate*	$P(R\|V)$	$P(R\|Q)$	$P(R\|C)$	$P(R\|C,D)$	*Precise*	*Recall*
peopleandnetworks-001	0.205	0.246	1.182	0.084	0.074	0.126	0.167	0.267	0.556
arthistory-001	0.289	0.335	1.125	0.102	0.074	0.109	0.188	0.453	0.190
dsalgo-001	0.177	0.355	5.961	0.061	0.082	0.075	0.038	0.182	0.540
pkuic-001	0.282	0.444	-0.649	0.077	0.088	0.117	0.151	0.328	0.545
aoo-001	0.247	0.328	1.446	0.090	0.056	0.071	0.042	0.351	0.583
bdsalgo-001	0.210	0.473	0.401	0.110	0.047	0.047	0.054	0.286	0.866
criminallaw-001	0.246	0.326	1.524	0.060	0.067	-	-	0.504	0.793
pkupop-001	0.283	0.428	1.122	0.095	0.091	0.126	0.212	0.356	0.596
chemistry-001	0.082	0.367	1.706	0.050	0.076	0.078	0.079	0.207	1.000
chemistry-002	0.413	0.494	0.707	0.056	0.042	0.071	0.036	0.362	0.696
pkubioinfo-001	0.260	0.332	-0.963	0.097	0.061	0.075	0.061	0.284	0.713
pkubioinfo-002	0.200	0.445	0.282	0.029	0.035	0.028	0.035	0.210	0.706

- **Jump-Random-Walk** (JRW), for representative students. It co-ranks the influence of both forum students and keywords meanwhile in G.

In order to compare with superposter, we set the same metric that a student is called a representative student when she is within top 5% of the rank list. Note that other alternative metrics, such as the threshold of an absolute number, are also feasible. The parameters used in JRW are $\beta = 0.85$, $\lambda = 0.5$ and $\epsilon = 10^{-6}$. $\lambda = 0.2$ and $\lambda = 0.8$ are also tried, however the differences are tiny. We adopt Normalized Discounted Cumulated Gain (NDCG) [10] as the metric which is applicable for evaluating rankings' quality. We invited two human judges who both are experienced in MOOC forums. They give the influence of each top 5% student a score by reading all the contents of related threads. Each thread and post here are preprocessed to be anonymous and unordered. Score values include 0, 1, 2 and 3, which denotes strongly disagree, disagree, agree and strongly agree. Finally the two assessments are averaged.

Figure 3 shows the results of human assessment. JRW outperforms others among the majority of courses as well as PR, which suggests the necessity of building such a heterogeneous network for discovering representative students. If instructors would set a rule to incentivize representative students, JRW could also be more objective and fairer than simple rankings based on the quantity of behavior. Here is a phenomenon that students voted the most are not representative. This is maybe by reason that the majority of forum students are actually not used to voting the influential posts while unusual comments earn many. In addition, we carry out the convergence analysis of JRW algorithm. Figure 4 shows this algorithm can converge rapidly and satisfy the requirement of real-time computation in large-scale applications.

6. ANALYSIS OF REPRESENTATIVE STUDENTS

In this section, we would explore the characteristics of representative students in two aspects of behavior and performance. Then based on the model and algorithm proposed, we developed a web service which can help instructors supervise not only the behavior and performance of each student, but also their relative position compared with the average level of the whole class. This service could be competent for instructors to gain feedback of the teaching quality.

6.1 Behavior and Performance

Firstly, we analyze the difference of behaviors between representative and non-representative students from a statistic view. Table 5 shows the proportions of various behavior of representative students to the whole forum students per course. The column of Forum Behavior contains three indicators, among which $P(R|T)$ and $P(R|P)$ reflect the degree of representative students' participation in forums. *Over Rate* indicates if the value is over zero, it means representative students' threads are more popular than the average, and vice versa. The values of the three indicators suggest in most forums representative students' participation is relatively high considering their low ratio, only 5%, and their threads are more popular. In other words, the result here manifests threads of representative students initiate the majority of discussions, not counting in the possible sub-discussions initiated by them within a thread.

The column of Learning Behavior shows the behavior of watching video and submitting quiz by representative students. The values of the two indicators, $P(R|V)$ and $P(R|Q)$, suggest the degree of learning behavior of representative students is relatively low compared with their participation, but still larger than 5%. So we can infer that representative students' learning behavior is just above the average. This also suggests their motivation is positive by judging from the value of $P(R|Q)$ which is related to the final certificate.

The column of Instructor demonstrates the necessity of preferentially answering the threads of representative students. *Precise* suggests instructors spent almost one third energy on answering representative students' questions, while *Recall* suggests instructors have answered about two third, up to overall, threads initiated by

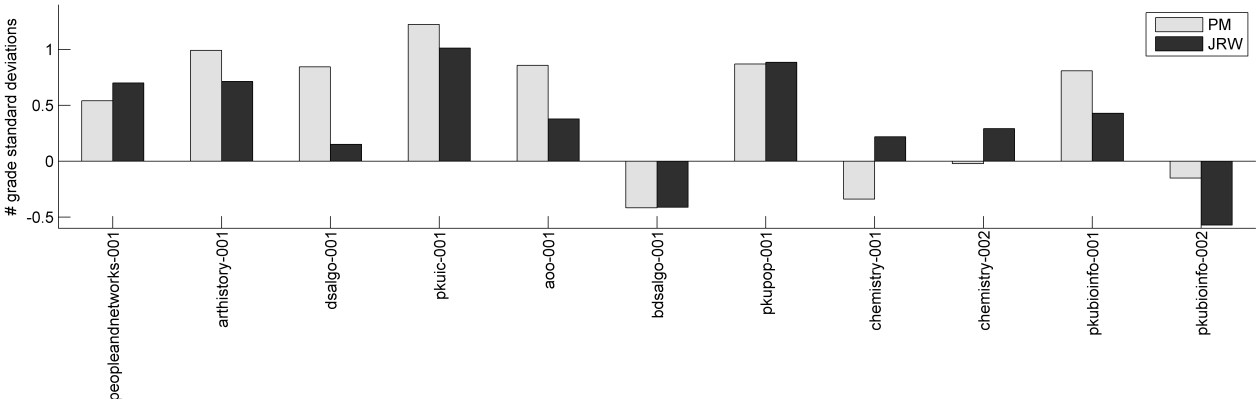

Figure 6: # of standard deviations of representative students outperforming non-representative students on grades per course, comparing with superposters by quantity.

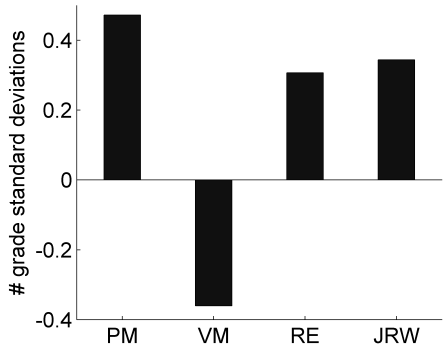

Figure 5: # of standard deviations of representative students outperforming non-representative students on grades averaged over all courses.

representative students. The historical records explain it is necessary for instructors to discover the representative students and their posts, since the range and time cost of choosing which post to reply from all are both reduced. The indicator of *Over Rate* also implies preferentially answering the threads of representative students means more audience would be indirectly beneficial, without actually having a response by the instructor.

Then we would analyze the performance of representative students in the forums. Still in Table 5, the column of Performance denotes the proportions of certificated representative students. $P(R|C)$ and $P(R|C,D)$ are indicators of the passed and the excellent representative students respectively. The values indicate representative students have the higher proportion among the excellent students than the passed students in most courses. However it is potential to improve the proportion of passing rate considering the large forum participation and positive motivation of representative students. So they are worthy being paid extra attention by instructors.

Figure 5 shows the standard deviations, that are averaged z-score grades, to illustrate whether representative students' averaged grade outperforms that of non-representative students among all courses, comparing four different ranking metrics. Superposter by quantity (PM), superposter by reputation (RE) and representative students by JRW (JRW) outperform their peers. However, the score of JRW is lower than that of PM. This may suggest representative students' performance is better than the peers, but not the group with best scores, and the top 5% students who post the most have the higher average score.

From the perspective of each course, representative students' performances are various. Figure 6 exhibits the same standard deviations per course. We can see representative students do not outperform their peers in some courses. Superposter and representative students almost show the consistent trends except for General Chemistry. Representative students' grade is lower than that of superposters by quantity in most courses, which also suggests representative students may have better performance above the average but not the best. This phenomenon could be explained that maybe similar to off-line class, representative students hard to master course content would involve more questions and need more instructions, while superposters by quantity are ones good at the course and always answer questions. So representative students are characterised by large participation of discussions, moderate learning behavior, and above-average performance but not the best.

6.2 Visualization Tool for Instructor

With the various forms of data, an open-and-shut visualization tool could be helpful for instructor to evaluate representative students and supervise their behavior. In order to apply the model proposed in previous sections to an actual function, we scale the final ranking scores to 0-100 as an index score, and developed such a web service whose interface looks as Figure 7.

Here we present the typical usage scenario of the service. Instructors could choose which course to see (Figure 7 A). Surely we would add role and permission administration to protect privacy in the future while here is just the demo of use cases. Then instructors could choose to see how many top students, at most overall (Figure 7 B). Instructors can also select to see the representative students' behavior (Figure 7 C) or their post contents (Figure 7 D). In the main exhibition area (Figure 7 C) where is a table list, instructors can realize the top students' various behavior, including forum participation, learning behavior and performance, students' influence index, and role in the forum. If instructors select to see 'influential post', the main area would replaced by the post contents composed by representative students (Figure 7 D). We conceive that Figure 7 D should provide functions for instructors to re-

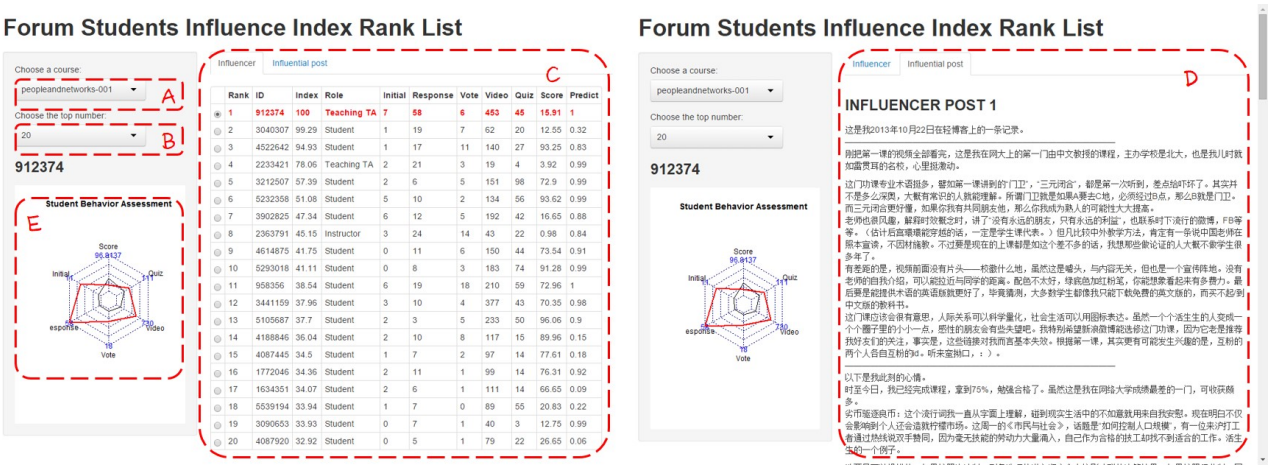

Figure 7: Web service interface

spond, rate, provide feedback and/or other post-related operations like those in the normal forum discussion settings in the future. Given the menu tab 'Influencer' selected, if instructors click the radio button ahead each record of the list, the behavior of corresponding student would also be presented in the radar chart (Figure 7 E). The radar chart displays six dimensions about students' behavior, that are quiz submission, video watching, vote, response, initiated thread, and final score. The scale of each dimension ranges from the minimum to the maximum of each class. Actually there are two closed hexagons on the radar chart. The fixed one in the middle denotes the average values in the whole class while the other, changed with trigger of radio click corresponding to each student, indicates the behavior of individual student. This radar chart can help instructors evaluate the behavior of each student comparing with the whole class under different dimensions.

In our observation and interview, this web service offers instructors the way to realize the class macroscopically and get feedback of main concerns in the forum promptly. Note that due to the rapid speed of our algorithm, this web service can real-timely refresh with changes of students' forum behavior.

7. CONCLUSION AND FUTURE WORK

In the MOOC forum settings, different participants may consider the influence as different definitions. We stand at the side of instructors and assume the influencers in MOOC forums are representative students who stimulate and attract much forum participation. They are actually characterized by lively engagement in forum discussions but unexpected learning behavior and performance, comparing with superposter. They are worthy being paid extra attention from instructors thereby to improve the course passing rate. Since they aggregate much discussion, they could be helpful to amplify instructors' answers and play the latent roles of knowledge propagation. Through representative students' influence, instructors can time-savingly realize the hot topics concerned by the most students. TAs' workload can be evaluated incidentally. In general, it is meaningful for instructors to preferentially read and answer representative students' threads.

In this paper, we leverage methods and algorithms of social network analysis to model MOOC forums in order to further understand the MOOC social learning settings and provide bases for instructors to intervene the social learning. This model has the advantages of fully utilizing social information and textual messages to identify and rank students' influence. Thus based on their behavior, performance and post contents, instructors may make measures to improve the teaching quality, better with that web service of visualization tool as an assistant.

Nevertheless, we have much future work to refine the discoveries in this paper. We would attempt other kinds of heterogeneous networks with more forum information and explore the effect of parameters. Some other random walk algorithms, such as HITS and topic based ones, would be more effective. Furthermore, by integrating our visualization tool into a practical platform, whether the amplification of knowledge propagation via representative students is effective and whether the teaching quality could be promoted still need to be verified through subsequent courses specifically designed in the future.

8. ACKNOWLEDGMENTS

This research was supported in part by 973 Program with Grants No.2014CB340405, NSFC with Grants No.61272340, No.61472013 and No.61370054.

9. REFERENCES

[1] P. Adamopoulos. What makes a great mooc? an interdisciplinary analysis of student retention in online courses. In *Proceedings of the 34th International Conference on Information Systems, ICIS '14*, 2014.

[2] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, pages 687–698. ACM Press, 2014.

[3] S. Bhatia and P. Mitra. Adopting inference networks for online thread retrieval. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI '10*, pages 1300–1305. AAAI Press, 2010.

[4] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, 8(1):13–25, 2013.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th*

International Conference on World Wide Web, WWW '1998, pages 107–117. Elsevier Science Publishers, 1998.

[6] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong. Learning about social learning in MOOCs: From statistical analysis to generative model. *IEEE Transactions on Learning Technologies*, 7(4):346–359, 2014.

[7] W. Cade, N. Dowell, A. Graesser, Y. Tausczik, and J. Pennebaker. Modeling student socioaffective responses to group interactions in a collaborative online chat environment. In *Proceedings of the 7th International Conference on Educational Data Mining, EDM '14*, pages 399–400. Chapman & Hall/CRC Press, 2014.

[8] HarvardX and MITx: The first year of open online courses, Fall 2012-Summer 2013. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263`.

[9] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders. Superposter behavior in mooc forums. In *Proceedings of the first ACM Conference on Learning @ Scale Conference, L@S '14*, pages 117–126. ACM Press, 2014.

[10] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '00*, pages 41–48. ACM Press, 2000.

[11] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, pages 1298–1306. ACM Press, 2011.

[12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03*, pages 137–146. ACM Press, 2003.

[13] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge, LAK '13*, pages 170–179. ACM Press, 2013.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM(JACM)*, 46(5):604–632, 1999.

[15] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pages 199–208. ACM Press, 2010.

[16] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and J. X. Yu. Influence maximization over large-scale social networks: A bounded linear approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*, pages 171–180. ACM Press, 2014.

[17] Q. Meng and P. J. Kennedy. Discovering influential authors in heterogeneous academic networks by a co-ranking method. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM '13*, pages 1029–1036. ACM Press, 2013.

[18] T. Schellens and M. Valcke. Fostering knowledge construction in university students through asynchronous discussion groups. *Computers & Education*, 46(4):349–370, 2006.

[19] A. Singh, D. P, and D. Raghu. Retrieving similar discussion forum threads: A structure based approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, pages 135–144. ACM Press, 2012.

[20] G. S. Stump, J. DeBoer, J. Whittinghill, and L. Breslow. Development of a framework to classify mooc discussion forum posts: Methodology and challenges. In *Workshop on Data Driven Education, Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013.

[21] H. Wang, C. Wang, C. Zhai, and J. Han. Learning online discussion structures by conditional random fields. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11*, pages 435–444. ACM Press, 2011.

[22] M. Wen, D. Yang, and C. Rose. Sentiment analysis in mooc discussion forums: What does it tell us? In *Proceedings of the 7th International Conference on Educational Data Mining, EDM '14*, pages 130–137. Chapman & Hall/CRC Press, 2014.

[23] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03*, pages 266–275. ACM Press, 2003.

[24] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang. Pagerank with priors: An influence propagation perspective. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI '13*, pages 2740–2746. AAAI Press, 2013.

[25] D. Yang, M. Piergallini, I. Howley, and C. Rose. Forum thread recommendation for massive open online courses. In *Proceedings of the 7th International Conference on Educational Data Mining, EDM '14*, pages 257–260. Chapman & Hall/CRC Press, 2014.

[26] R. Yeniterzi and J. Callan. Analyzing bias in cqa-based expert finding test sets. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14*, pages 967–970. ACM Press, 2014.

[27] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han. Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14*, pages 283–292. ACM Press, 2014.

[28] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 7th IEEE International Conference on Data Mining, ICDM '07*, pages 739–744. IEEE Press, 2007.

[29] G. Zhou, S. Lai, K. Liu, and J. Zhao. Topic-sensitive probabilistic model for expert finding in question answer communities. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 1662–1666. ACM Press, 2012.

[30] H. Zhu, H. Cao, H. Xiong, E. Chen, and J. Tian. Towards expert finding by leveraging relevant categories in authority ranking. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, pages 2221–2224. ACM Press, 2011.