Advances in Engineering Research, volume 150

4th International Conference on Machinery, Materials and Computer (MACMC 2017)

Analysis of Online MOOC Learning Network Structure based on the Social Network

Hong LIU[1,a], Xiaojun LI[2,b*] and Biwei LI[2,c]

[1]Computer and Information Engineering College, Zhejiang Gongshang University, Hangzhou, 310018, China

[2]School of Management and E-Business, Zhejiang Gongshang University, Hangzhou, 310018, China

[a]email:LLH@mail.zjgsu.edu.cn, [b]email:lixj@mail.zjgsu.edu.cn, [c]email:libiwei@mail.zjgsu.edu.cn, [*]corresponding author

Keywords: MOOC, Learner, Social Network, Course, Clustering, Community

Abstract. Through the online learning platform, learners form a virtual learning network. With such the learning network, learners can learn relevant knowledge and communicate with other learners about study notes, which demonstrate the similar characteristics of network with strong or weak relations. From viewpoints of the course and the learner, this paper applied the social network analysis method to study the factors, such as the degree of strength, the network structure and the willingness of information transmission, among the individuals on the MOOC learning platform. The paper analyzed the features of density, transmission, clustering and community based on the online MOOC learning data, whose experimental results show that online course based social network has some characteristics of high network density, medium information transmission and common community. For the learner social network constructed for a specific course, it is found that there is less communication about learners' identities and occupations and stronger clustering.

Introduction

With the rapid development of information technology and network technology, online learning platform and database technology has been vigorous developed and applied in the field of online education, and the characteristics of autonomous, sharing and open made it an excellent complementary to the traditional education. The MOOC, a rapid development in recent years in the world, is a new type of education for learners and provides more opportunities for them to pursue college courses. Based on the MOOC platform, many learners could form a virtual learning network through online course discussion and knowledge sharing. Based on this network, learners can learn relevant information and exchange and share the learning experience, which makes the network have the characteristics of network relationship, such as the strength of relations.

During the MOOC study, the learners have produced a vast amount of data related to their learning behavior, such as the time of learning, the learning content, the evaluation of the course, etc., which can reflect the learning habits and styles from the deep level. However, in the learning process, a variety of factors may affect the collaboration and communication of online learners, such as the learner's identity, positive degree, satisfaction and so on. Facing the large amount of data generated from the learning platform, how to analyze these data effectively and study the behavior pattern of online learning so as to guide and evaluate the learners' learning and effects has become a hot issue.

At present, in the field of online learning behavior analysis, the domestic research focuses on two dimensions: data index and technical means. For data index, Xu Bin [1] classified learners according to their different behaviors and established an interaction network among learners after adopting social network analysis method to analyze the structure characteristics of the user node. Zeng Xiangyue [2] investigated and analyzed the online learning behavior of distance learners, such as online learning time and learning resource utilization rate. After analyzing the behavior characteristics of online learners' interaction activities, Jin Li [3] constructed an online learning

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

behavior model. Yang Jinlai [4] proposed the Who-Do-What network behavior information model. Li Feng [5] analyzed the learners' learning behavior of large-scale network courses. Jiang J and etc. [6] mined the potential friendship among learners through analyzing the online interactive activities. Griinewald F et al [7] found the social collaboration among the communication and interaction on the MOOC platform. For technical means, Zhou Yan [8] modeled and analyzed the network learning behavior for college students and constructed a causal model with nine latent variables based on the Theory of Reasoned Action and Technology Acceptance Model (TAM). Li Dejiang et al [9] designed and implemented a distributed online learning behavior statistics system, including the data acquisition of distributed clients and the transmission to the server statistics. Liao Jing [10] designed a scheme of data acquisition and behavior analysis of online data stream and summarized the online learning behavior. Pan Lei et al [11] proposed an active model to analyze user behavior to obtain behavior patterns, which could be used to determine behavior tendency and detect abnormal behavior. Jing Jiang et al. [12] analyzed the potential interaction among users of a specific course on the web site, besides online discussion.

Foreign scholars focused on three aspects of online learning: tracking and recording of online learning behavior, learner needs and online learning environment, and the relationship between online learning behavior and learning performance. Apostolos et al [13] analyzed the evaluation and experience of MOOCs through Twitter and concluded the learners' feelings and needs in MOOCs learning to help better design the MOOCs platform. Franka et al [14] pointed out the high social collaboration characteristics of communication among middle school students on the MOOCs platform. Anderson [15] attempted to categorize participants by analyzing user motivations for MOOCs learning. Yang and Tsai [16] used questionnaires to collect information and explored the learning environment preferences and learners' beliefs in online learning. Criatiobal et al. [17] proposed an advanced architecture to improve the efficiency of web mining in personalized learning system. For poor quality of online learning, Alkhattabi et al. [18] used web data mining technology to establish an evaluation model of online learning quality.

In this paper, the data of MOOC online learning platform is used and the social network analysis method is adopted to study the characteristics of social network of online learning, including network density, transitivity, clustering and community characteristics. Factors are also studied to understand the degree of strength between individuals, the network structure and the willingness of information transmission.

Analysis of Online MOOC Learning Network Structure

Data Collection and Processing. We use the network robot to crawl the course address for the learning record on the MOOC platform.The original crawled online learning information (such as course name, category, evaluation, and etc.) may contain some noise, such as spelling errors, content repetition and information asymmetry, so preprocessing is required, such as web page cleaning to remove pages with noise and automatic filtering to delete the repeated resources to save storage space. After preprocessing, the information is transformed into relational data structures and stored in local database. The specific data structure used for this experiment includes: course information (course categories, course ID, course name in English, course name in Chinese, the number of students, course interest measure), user information (user ID, user name, level of education, the number of favorites) and learning information (user ID, course ID, user URL).

Definition of Social Network. Based on MOOC learning platform, if the learner is regarded as a node and the relationship between course and learner is taken as an edge, then a network structure can be formed among learners and learning courses. In the same way, if the course is taken as a node and the learner's relationship between two courses is taken as an edge, then a network structure could be formed between course and learners.

Definition 1: Course-Course Social Network, referred to as CCSN, with the structure: CCSN= (V, E), where: V= (C_1, C_2, C_3, ..., C_n) is a collection of course nodes and E is a collection of undirected edges. CCSN represents the relations of learners who have taken two arbitrary courses, thus (C_i, C_j)

indicates that the learner L_i takes the courses C_i and C_j at the same time. The adjacency matrix of CCSN describes the connection between courses and courses:

$$\varphi_{ij} = \begin{Bmatrix} w, & (C_i, C_j) \in E \\ 0, & otherwise \end{Bmatrix} \tag{1}$$

Where φ represents an n*n matrix, n=|V| is the number of network nodes and m=|E| is the number of network edges. If $\varphi_{ij} = w$, then there are w learners taking both courses C_i and C_j.

Definition 2: Learner-Learner Social network, referred to as LLSN, with the structure: LLSN= (V', E'), where: V' = (L_1, L_2, L_3, ..., L_n) is a collection of learner nodes and E is a collection of undirected edges. LLSN indicates the relations of any two learners taking the same course, thus (L_i, L_j) represents learners L_i and L_j taking course C_i at the same time. The adjacency matrix of LLSN describes the connection between learners and learners:

$$\varphi'_{ij} = \begin{Bmatrix} w', & (L_i, L_j) \in E' \\ 0, & otherwise \end{Bmatrix} \tag{2}$$

Where φ' indicates an N*N matrix and N'=|V'| is the number of network nodes. If $\varphi'_{ij} = w'$, then both learners L_i and L_j take w' courses.

Characteristics of MOOC Learning Network. In this paper, the indexes of social network are selected, such as network density, transitivity, clustering and community characteristics, which are used as quantitative indicators for structural analysis of MOOC learning network.

(1) Density: it shows the degree of closeness between nodes in a network, which could be used to measure the degree of connectivity between nodes in a social network. The value of density is between 0 and 1, and the closer the value is to 1, the closer the relationship represents. But the greater the density, the greater the influence of the network on the attitude and behavior of the nodes. Therefore, a moderately dense social network facilitates communication between nodes.

(2)Transitivity it shows the performance of information communication between nodes in a network, which is represented by the degree and the clustering coefficient of each node in the network. The equation is:

$$T(G)=3*\frac{\text{the number of triangles in G}}{\text{the number of trituples}} \tag{3}$$

where the number of triples refers to the number of edges with common nodes.

(3)Clustering coefficient: it shows the degree of clustering and the cohesive tendency of the network. For example, the circle of acquaintances or friends in the social network, in which every member knows other members. The equation is:

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{4}$$

where $C_i = \frac{e_i}{k_i(k_i-1)}$, k_i is the degree of node i and e_i is the number of edges between node i and adjacent nodes.

(4) Compatibility: it shows the description of the association between nodes with small and large degree.

(5) Reachability: it shows the ability of any node in the network to establish an effective connection with other nodes, which is represented by the network average shortest path.

(6)Centrality: it shows the importance of the node in the network. Degree centrality and betweeness centrality are used to describe the characteristics of the network centrality. The former one represents the importance of the node, describing its position within the social network (inherent right); while the latter refers to the importance of the node from the angle of overall integration.

(7) Community: it shows nodes in the same set have relatively strong, direct and close connections, while nodes of different sets are not so related. So the communication of nodes in the same set is smooth, and if there are a certain number of sets, then the communication between nodes of different set would become less smooth.

Community Division Method based on Infiltration Algorithm (Clique). Considering the need to cluster large data space, since the clustering results are not sensitive to the data input order and the shape of clustering is not known in advance, there is no need to assume any canonical data distribution. This paper uses the infiltration algorithm (Clique) to perform community division for the constructed social network. The basic idea of the algorithm is to divide the data space into several grid cells. The number of data objects in each cell is the density of the cell. When the density of a specific cell is greater than the given threshold, the cell is regarded as dense, and the final cluster is the maximum connected interval of adjacent dense cells. For graph G, if there is a complete subgraph (there are edges between any two nodes) and the number of nodes is k, then the complete subgraph is called a k-clique. Furthermore, if there are k-1 common nodes between two k-cliques, then the two cliques are called adjacent. A string of adjacent cliques that constitutes the largest collection is known as the community.

The details of Clique algorithm applied to CCSN and LLSN are described in the following:

```
Algorithm：Clique algorithm based on CCSN(LLSN)
Input：Connect matrix $\varphi$(or $\varphi'$), k(dimension)
Output：the number of communities n and the list of nodes in each community
1. Locate all the one-dimension dense regions of each attribute
2. k=2
3. repeat
4.     Generate all the candidate k-dimension dense units from k-1 dimension dense units
5.     Delete all the units smaller than the threshold
6.     k=k+1
7. until all k-dimension dense units no longer exists
8. Through selecting all adjacent and high-density units, clustering is discovered and all
   the nodes in each cluster are stored into the list.
9. Output the number of communities n (the number of clustering) and all node indexes
   for each community from the list
```

Experimental Results and Analysis

In this paper, the online learning record of the MOOC platform is crawled to become the research target. The total number of learners is 5638 and the number of courses is 1791. Python3.6 is used as the analysis tool and SqlServer2008R is used for data storage.

CCSN Analysis. By processing the crawled data, we get that the number of nodes is 1791, the number of edges is 170156, whose weight w changes within [1,157]. Because of the obvious differences in weight, the constructed social network would have too many independent nodes, if the data is not properly filtered. For example, if nodes and edges with weights greater than or equal to 10 are used for construction, then the independent nodes (i.e., nodes that do not have associated relationship with other nodes) are nodes 712, 1347, 207 and 1769. So in the experiment, we select the weight greater than or equal to a threshold value nodes to construct the social network. After many experiments, this paper chooses nodes with w greater than or equal to 20 to construct the social network. Both C_i and C_j courses that have been chosen by more than 20 learners are selected to construct the following network structure shown in Figure 1, where the number of nodes is 52 and edge number is 236. If the independent nodes are removed from the network, the whole network is interconnected, which is beneficial for subsequent analysis.

The distribution of the degree of CCSN network nodes is shown in Figure 2, which can be seen that network node degree is rather heterogeneous. A large degree indicates that this course is a popular course taken by more learners and vice versa.

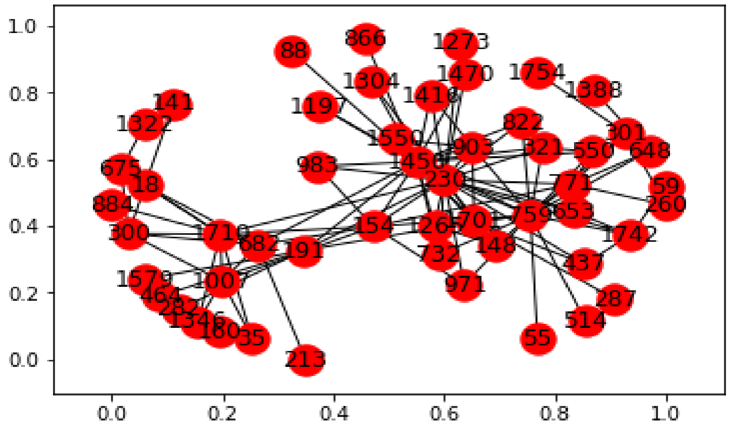


Fig.1. The CCSN structure with w>=20

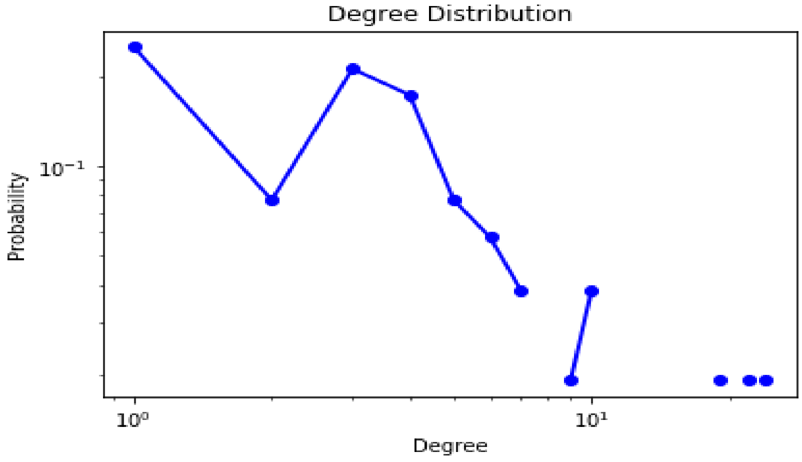
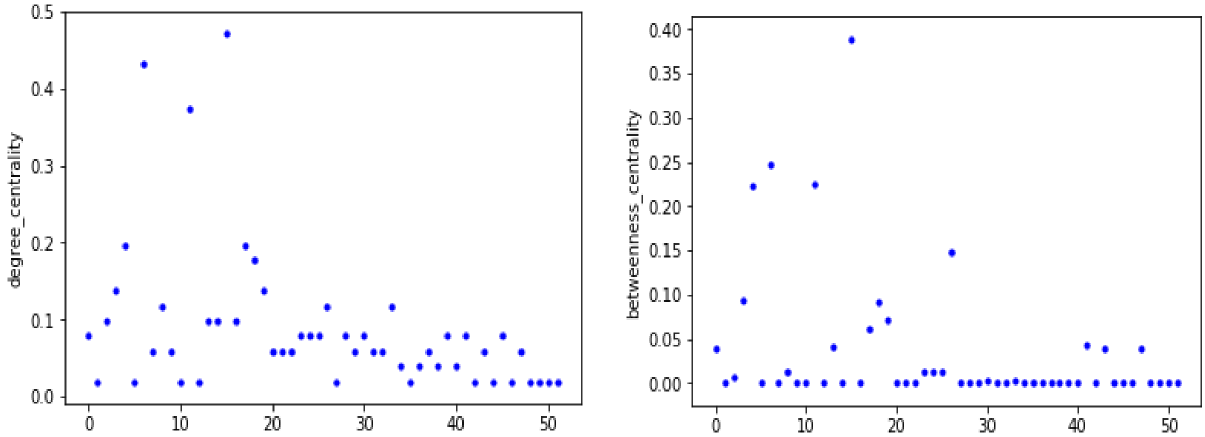


Fig.2. The distribution of degree of CCSN nodes

Figure 3 shows the distribution of CCSN nodes, and there exists some high centralized nodes (for course nodes 230, 1450 and 759, the degree coefficients are 0.4706, 0.4314 and 0.3725 and the betweenness coefficients are 0.3890, 0.2474 and 0.2257 respectively), indicating the importance of these courses in the network, namely the core course in the network. Since more people learn these courses, it can be considered as the preferred recommended courses. Overall, the higher the degree coefficient, the higher the betweenness coefficient, and vice versa. Of course there are some nodes, such as course 1701 with the degree coefficient 0.1961 (ranked fourth), its betweenness coefficient is only 0.0613 (ranked ninth). These nodes should also be paid attention to when courses are recommended.



(a) Degree centrality distribution (b) Betweenness centrality distribution

Fig.3. The centrality distribution of CCSN nodes

Other features of CCSN network are listed in Table 1.

Table 1 Other features of CCSN and LLSN

Indicator	CCSN Value	LLSN Value
Network density	0.089	0.006
transitivity	0.296	0.106
Clustering coefficient	0.519	0.861
Compatibility	-0.317	-0.609
Reachability	2.750	2.343

As shown in the table above, the CCSN network density is 0.089, indicating a loose structure and low relevance association of the courses. The transitivity is 0.296, indicating low willingness of the learners to exchange information of the course content, and the clustering coefficient is relatively large,0.519, indicating the existence of small groups, namely the possibility of a condensed subgroups. The compatibility is -0.317, indicating nodes with smaller degree are connected with nodes with higher degree, but the reachability is 2.750, which represents the distance for these two courses to be connected through learners.

The CCSN network is divided into four communities by the infiltration algorithm.

Frozenset1 ({18, 154, 682, 300, 1710, 1007}),

Frozenset2 ({464, 282, 160, 1579, 191}),

Frozenset3 ({1346, 35, 1710, 1007}),

Frozenset4 ({321, 771, 903, 1416, 971, 653, 1742, 1550, 148, 983, 154, 732, 1701, 230, 550, 1450, 1265, 822, 759, 191})

The courses of community 1 are mainly related to data processing and analysis. The courses of community 2 are mainly Chinese history. The courses of community 3 are mainly computer focused, and the number of courses in the community 4 is relatively huge. Among these four communities, nodes 191, 1710, 1007, 154 are the overlapping nodes, through which different communities are interconnected, so learners of two communities can communicate through these overlapping nodes.

LLSN Analysis. CCSN network analyzes the MOOC network structure from the perspective of course-course. To further study the relationship of learner-learner, this paper selected a course with moderate number of learners and the overlapping node between two communities of the course to construct the LLSN as the object of the study. For course number 1007, *the R language development course*, the number of LLSN nodes *n* is 131 and the number of edges is 112476. Figure 4 shows the degree distribution of LLSN nodes. It can be seen that the degree of node distribution in LLSN is not centralized, which is in line with the characteristics of the nodes selected by LLSN in this experiment. In the experiment, learners of course 1007 are selected and the connections do exist between the nodes in LLSN. The value of w is at least 1 and if any two learners have learned other similar courses, then w is bigger than 1.

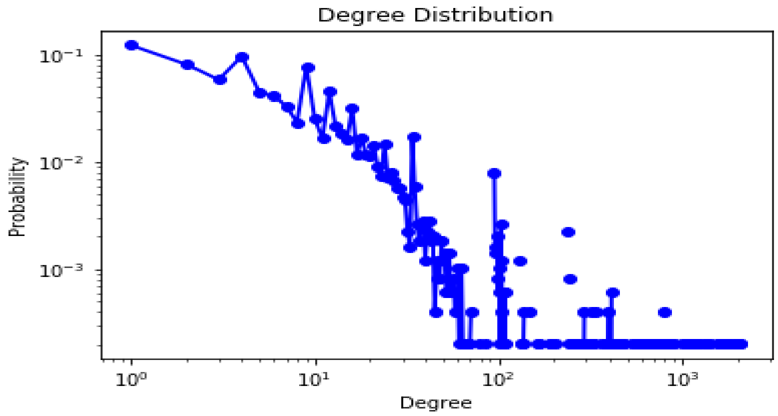
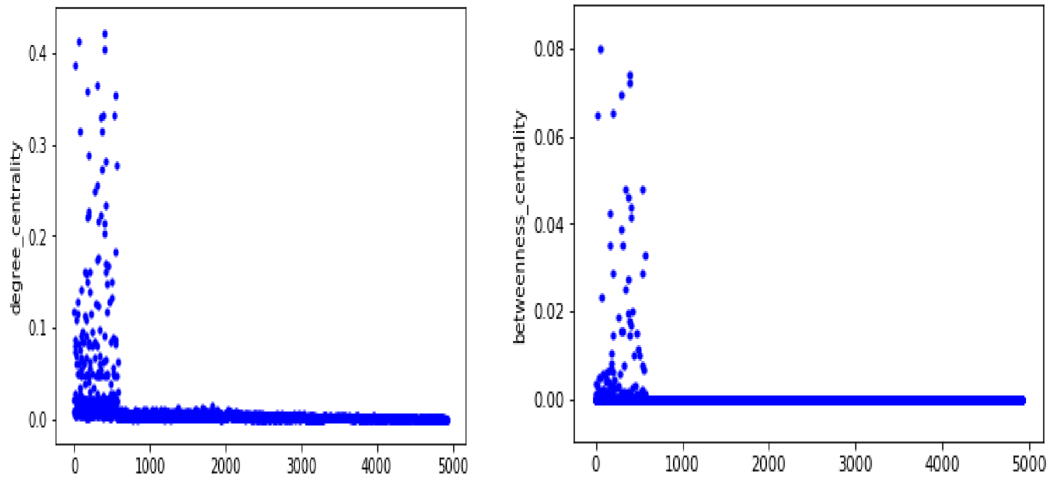


Fig.4. The distribution of degree of LLSN network

Figure 5 shows the central distribution of the LLSN nodes of course 1007, which is similar to that of CCSN. The nodes with low degree of centrality also have low betweeness coefficients, and vice versa. There are some nodes with high centrality, namely active members, act as the intermediary members and play a key role in propagating course learning. They communicate information and are conducive to the cooperation learning of other courses.



(a) Degree centrality distribution (b) Betweenness centrality distribution

Fig.5. The centrality distribution of LLSN nodes

Other characteristics of LLSN are shown in Table 1, it can be seen that the density of LLSN network is very low, the distribution of learners is very sparse and the information transmission of courses is not very high. Except for certain intermediary nodes, the agglomeration of the network is very good, which results in only one community when the Clique algorithm is applied to the LLSN network, in accordance with the data requirements. The compatibility value is smaller than that of CCSN, so is the reachability value. It shows that the distance between any two learners is relatively short for the same course. For all courses, two associated courses have a relatively long distance.

Conclusions

With the development of online learning platform, learners have generated a lot of data in the process of online learning. Through analyzing the data, the learners' interest and learning behavior can be explored, which is helpful to guide the study and recommend relevant course. In this paper, the social network analysis method is used to analyze the learning data on the MOOC platform. The CCSN network is constructed based on the course and the learning behavior is analyzed from the perspective of course. It is found that the network density is low, the information transmission is general and the network has obvious community characteristics. On the basis of CCSN network, the LLSN network was constructed for a specific course and the learners of the same course were analyzed. It is found that the communication was even less because of the differences of learners' identities, occupations and so on. Based on the above-mentioned work, we will consider the following research work: applying the CCSN network to obtain the relationship between courses and providing learning path recommendation for learners with integration of network density, accessibility and centrality.

Acknowledgement

In this paper, the research was sponsored by the Online Education Research Fund (All-pass Education) Project of MOE Research Center for Online Education (Project NO.2016YB144), and the National Social Sciences Funding Project (Project No. 17BTQ069, 14BTQ047).

References

[1] Xu Hongcai. Investigation and Research on Network Learning Behavior of College Students [J]. E-Education Research, 2005 (6): 61-63.

[2] Zeng Xiangyue, Yuan Songhe. Investigation and Analysis of Distance Learners' Network Learning Behavior [J]. China Journal of Distance Education, 2008 (4): 47-51.

[3] Jin Li. Design and development of online learning behavior and related data analysis [D]. Inner Mongolia Normal University, 2008.

[4] Yang Jinlai, Zhang Yixiang, Ding Rongtao. Research on Learning Behavior Monitoring Based on Web-based Learning Platform [J]. Computer Education, 2008 (11): 65-68.

[5] Li Feng, Li Jie, Zhao Changkuang, eta al. Analysis of learners' learning behavior of large-scale network course [J]. Computers education,2014,20(10):49-52.

[6] Jiang J,Wilson C, Wang X,et al. Understanding latent interactions in online social networks[J]. ACM Transactions on the Web,2013,7(4):18.

[7] Griinewald F,Meinel C,Totschning M,et al. Designing MOOCs for the support of multiple learning Styles[M]. Rerlin:Springer,2013:371-382.

[8] Zhou Yan. TRA and TAM based college student network learning behavior model [J]. China Audio-Visual Education, 2009 (11): 58-62.

[9] Li Dejiang, Zhang Peng, Tian Zhiying, Zhu Gehua, Ru Pengxin. Design and Implementation of Distributed Learning Behavior System[J]. Journal of Xinjiang Radio & TV University, 2010 (2): 5-9.

[10] Liao Jing, Zhang Hui. A flexible online learning behavior data acquisition and analyzing system [J]. Information and Computer (Theoretical Edition), 2011 (1): 85-86.

[11] Pan Lei, Zhu Hongxia. Research and Design of Network Access Behavior Analysis Model [J] .Computer & Modernization, 2011 (9): 140-143.

[12] Jing Jiang, Christo Wilson, Xiao Wang, et al. Understanding Latent Interactions in Online Social Networks[J]. ACMTrans. On the Web,2013, 7(4):18:1-18:39.

[13] Apostolos Koutropoulos, Sean C. Abajian, Inge deWaard, el . What Tweets Tell us About MOOC Participation [J]. International Journal of Emerging Technologies in Learning, 2014, 9(1): 8-21.

[14] Franka Griinewald, Christoph Meinel, Michael Totschnig et al. Designing MOOCs for the Support of Multiple Learning Styles[J]. Scaling up Learning for Sustained Impact(Lecture Notes in Compute Science),2013,8095:371-382.

[15] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec .Engaging with Massive Online Courses [C]. ACM International Conference on World Wide Web (WWW'2014)Seoul, Korea, 2014.

[16] Yang, F.Y., & Tsai, C.C. Investigating university student preferences and beliefs about learning in the web-based context[J]. Computers & Education,2007,50(4):1284-1303.

[17] Criatiobal,R., Sebaatian, V., & Amelia, Z. Applying Web usage mining for personalizing hyperlinks in web-based adaptive educational systems[J]. Computers & education, 2009, 53(3):828-840.

[18] Alkhattabi, M., Neagu, D., &Cullen,A. Assessing information quality of e-learning systems: A web mining approach[J]. Computers in Human Behavior,2011,27(2):862-873.