Interactive Discourse Analysis Based on the Forum Text Mining in Cloud Classroom

Zhifeng Wang, Rong Zhao, Yanli Xu, Xiangyong Li, Mingzhang Zuo, and Junmin Ye

Abstract—**The MOOCs have risen in online learning and they develop and change constantly. At present, the universities not only join the MOOCs platforms but also build their own online learning platforms. With applying the online learning platforms, such as cloud classrooms, for hybrid teaching, a bunch of learning data have generated on those platforms, but utilization of such data for developing learning effect need to be pained more attention. Based on the learning characteristics of forums and the needs of forum learning assessment, this paper builds an interactive discourse analysis mode based on the cloud classroom platform of Central China Normal University, and a real course data is applied to conduct empirical research. Research results show that this mode can effectively describe forum learning behavior of student and can assist teachers in screening special students to help them make the appropriate interventions.**

Index Terms—**Learning analytics, interactive discourse, text mining, cloud classroom.**

I. Introduction

MOOCs (MOOCs), which are originated from online distance education and online video courses, firstly appeared in 2008. The developments of MOOCs are constantly in the ascendant. Nowadays, the colleges and universities are all engaged in MOOCs platforms, such as Coursera, Udacity, edX [1]. Many colleges and universities use the idea of online education to establish their own online learning platform, which are followed by stable audiences. Teachers use the online learning platforms to assist their teachings and achieve multiple teaching goals.

Saitta [2] pointed out that classrooms with interactive dialogues can significantly enhance students' intrinsic motivations, interests and ability to learn and promote classroom vitality. With the rise of courses taken through MOOCs, the analysis of interactive discourse should not only be confined to the offline classroom, but also online learning platforms will produce more and more interactive discourse data, including text, pictures, voice, video and so on [3]. Among them, the text is the most common interactive method used by teachers and students through the posts in forums, which is a more realistic reflections of the learners' motivations, cognitive developments, emotional attitudes, and learning experiences. Text exchange through online forums is an important aid for online learning and blended learning. Especially in the online learning platform established by colleges and universities, the forum is a commonly regarded as an important auxiliary teaching tool by teachers in blended learning. The students' important learning contents and learning behaviors can be extracted through the forums of online learning platform. Through the excavation and description of students' learning behaviors on forums, they can effectively reflect the learning status of students, which can help teachers to screen special speaking students, and monitor and correct the overall behavior of students. Current online learning platform can achieve the learning situation feedback, including learning progress, learning duration, clicks and so on. Normally, teachers learn about students' performance by controlling topics and browsing posts. Studies have found that the quality of interaction at the initial stage of a course has a continual and significant effect on the level of interaction at the end of the course [4]. Therefore, improving the interaction levels of the online learning forums is an important prerequisite to improve the quality of learning and teaching.

II. Related Research

The concept of discourse analysis was firstly applied to the field of pedagogy [5]. With analyzing the interactive discourse between people in a particular learning situation, the discourse analysis is used to explore the semantic content, identify organizational features of discourse, and understand knowledge structure and learning patterns. In other words, although the understanding of discourse content is the main purpose of forum learning analysis, we cannot neglect the study of the relationship between organizations and connections among discourses.

The research of interactive discourse analysis can be divided into three categories: interactive discourse classification [6]-[8], interactive features analysis of discourse [9], [10], and discourse interaction design [10]-[12]. These three types of research are not mutually exclusive categories, but rather staggered. The research of interactive discourse classification is mostly motivated by the perspective of educational research. The study of interactive features analysis is involved in the technical perspective of learning analytics, which focus on data analysis. The discourse interaction design is from the view of system methodology, which pay attention to the actual effect of the

Manuscript received August 27, 2018; revised November 12, 2018. This work was supported by the National Natural Science Foundation of China (No.61501199), Hubei Natural Science Foundation (No. 2017CFB683), self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No. CCNU18QN021), and the National Science and Technology Ministry (No. 2015BAK33B02).

Zhifeng Wang, Rong Zhao, Yanli Xu, Xiangyong Li, and Mingzhang Zuo are with the School of Educational Information Technology, Central China Normal University, Wuhan, Hubei 430079 China (e-mail: zfwang@mail.ccnu.edu.cn).

Junmin Ye is with the School of Computer, Central China Normal University, Wuhan, Hubei 430079, China.

doi: 10.18178/ijiet.2019.9.3.1195

method.

In the first research category, the interactive discourse is divided into three types in [6]: operation interaction, information interaction and concept interaction. These three interactions cover the most important and fundamental aspects of forum learning discourse analysis. In addition, some scholars divided learning analysis tools into five categories: content analysis, network analysis, behavior analysis, ability analysis and comprehensiveness analysis. Among them, the most relevant is content analysis and network analysis tools.

For interactive features analysis of discourse, some studies focus on learners' click behaviors, learners' interaction and cooperation. However, none of them builds an analytical mode that can be used for forums discourse analysis. The most important thing in this kind of research is the mode construction and algorithm realization of the clustering and theme mining, ignoring the pedagogical meaning represented by each index. Finally, it can be observed that different discourse categories have different depth affective tendencies. This model has the potential to automatically analyze and deeply interpret the content and structure of online discussions of conversations to better provide learners with an adaptive experience.

The research on online learning interaction design mainly focuses on the following aspects: theoretical basic research, relationship between interaction degree and learning effect, measurement and evaluation of interaction quality, interaction tool research and design, factors affect the interaction in interactive environment. The study of learning behavior is more concerned about the forum to discuss the theme of discovery and mining.

To sum up, many existing studies provide some reference ideas for building forum learning behavior analysis, such as the selection of analysis indicators, available methods and tools. Among the vast data generated by the Internet, the texts can be said to be the most difficult data to be collected but the most difficult to be analyzed because the understanding of texts is a subject to be further overcome in the current research work of nature language understanding. Under the premise of the current development of text analysis technology, how to effectively use the existing analytical techniques and data to analyze the forum text data generated by students in online courses to promote the development of student learning and teacher teaching is an urgent problem to be studied.

In this paper, the theme of forum learning behavior analysis, combined with the learning characteristics of forums and the needs of forum learning assessment, systematically researches the forum analysis tools and forum learning analysis cases and constructs an interactive discourse analysis based on a university cloud platform and uses one curriculum data for empirical research.

III. Construction of Interactive Discourse Analysis Mode Based on Text Interaction

A. *Analysis Mode Construction*

Through the research on the current forum evaluation system, text analysis methods and tools, the teaching evaluation index as the main structure, combined with the current technology can achieve the target description, we build a forum based on a university cloud interactive discourse analysis mode.

In the interactive discourse analysis, the three types of data should be timely combined. The interactive discourse analysis mode proposed in this study is based on the former two types of data-based automated machine analysis, supplemented by the third type of data analysis to help the analysts even more accurately grasp the accuracy of the analysis.

The proposed interactive discourse analysis mode based on forum text interaction includes three types of factors. The first category is the internal information of the text. The semantic analysis needs to find the summary of the text, emotions, categories, themes and keywords. The second category is the text of the external information, such as the text of the release time, the number of words in the text, the text of the publisher, the text posted reply relations. This information can generally be extracted directly from the forum background data. In the modeling and selection analysis dimensions, we cannot consider variables with low relevance such as gender, as it has been explicitly mentioned in previous studies that such variables have a particularly small impact on the learning effect. The third category is the text of the extension of information, such as the meaning of the text. Such textual information requires manual analysis or more advanced semantic analysis methods, and the current computerized semantic analysis has not yet reached the desired level.

B. *Analysis Methods and Tools*

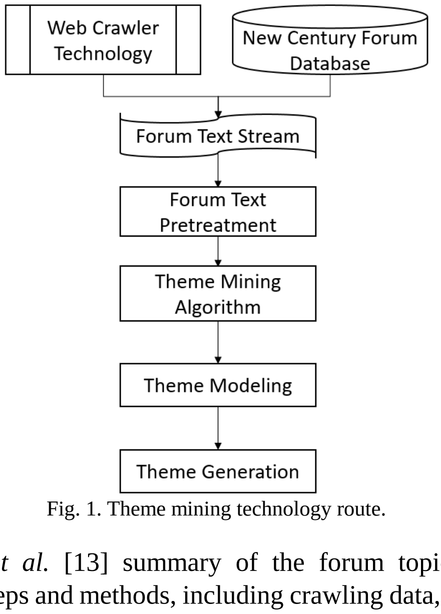


Fig. 1. Theme mining technology route.

Chen *et al.* [13] summary of the forum topic mining general steps and methods, including crawling data, access to text streams, text preprocessing, theme mining algorithms, theme modeling and theme generation. The realization of the path shown in Fig. 1.

The mode constructed in this paper includes all the methods of topic mining. In addition, through the methods of social network analysis and sentiment analysis, the rich learning information contained in the text is excavated at different levels. Through the cross-understanding of different levels of information, the implicit meaning of the learner's text expression is extracted, the learner's opinions and ideas are evaluated, the learner is understood, and the intervention and reasoning in the learning of the learner's forum are promoted, so as to promote their realization Knowledge-driven deep processing. Technical route shown in Fig. 2 below.

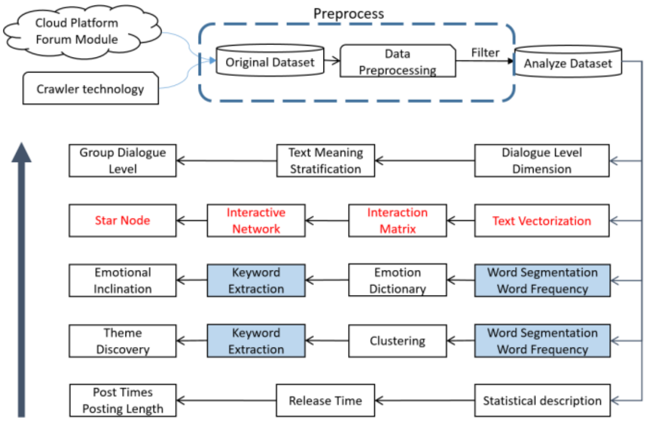


Fig. 2. text-based interactive discourse analysis mode technical route.

1) *Data collecting*

Data collecting refers to the use of a program to automatically collect useful information in the network, usually used to collect text messages. Web pages contain a lot of information such as texts, HTML tags, script scripts, etc. To collect useful text information, we need to parse the web page to remove unwanted markings and identify disturbed text information, and then automatically write the text information to be collected in the local database. Currently there are more mature integrated crawler tools for direct use, such as goo seeker. This is a python language based tool. The operation steps shown in Fig. 3 below.

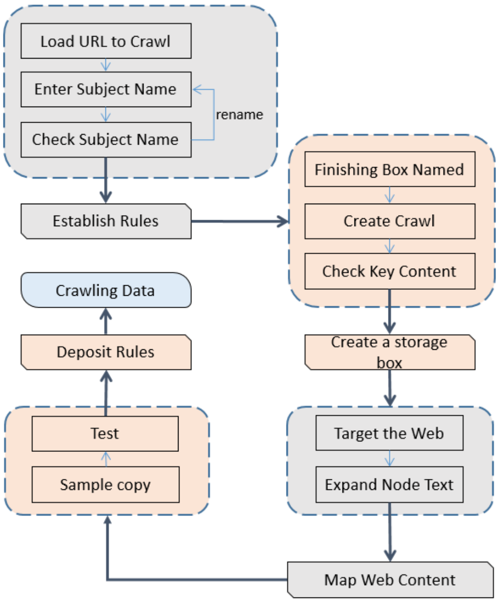


Fig. 3. Data crawling technology route.

In general, a crawler uses a URL or URLs of some kind as a seed to learn some sort of crawl rules through training programs, which can then be used directly as a method. In the official crawl process, usually from a page to get and record all the pages of the link and sort the link, and then in the new page to save and extract the link action, and so on, until all the mining the web page specified in the mission was visited. Therefore, in the specific operation, to determine the scope of crawling.

2) *Text preprocessing*

The purpose of forum text preprocessing is to filter out invalid data in the original forum data, converting the text into data objects that facilitate the computer's processing of the calculations [13]. Text preprocessing generally includes the steps of word segmentation, stop words, word frequency statistics, word co-occurrence and text vectorization. The ICTClAS word segmentation system used in this paper can meet the functions of word segmentation, word frequency statistics and word co-occurrence statistics in this study.

The main idea of the ICTClAS participle system is to classify words by CHMM (Layered Markov Mode). By stratification, the accuracy of word segmentation is increased and the efficiency of word segmentation is guaranteed. There are five layers, as shown in Fig. 4. The basic step is to conduct atomic segmentation first, and then on the basis of the N-shortest path rough segmentation, find the first N most consistent segmentation results, generate binary vocabulary, and then generate the word segmentation results, followed by POS tagging and complete the main word segmentation steps.

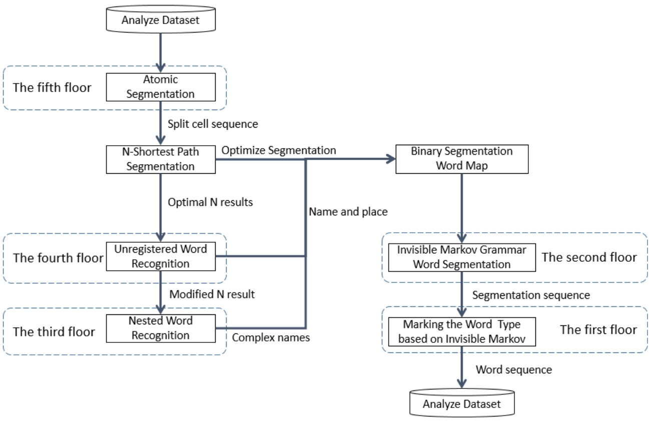


Fig. 4. Text preprocessing framework based on CHMM.

3) *Theme mining*

The preprocessed data is basically in line with the machine-handling standards, but the preprocessed data is still sprinkled with a few loose, unmeasurable data content. Theme mining technology is the separation of these free data, extract the relevant data to form a specific topic. Due to the particularity of the forum structure, the key elements such as the title, the post, the user and the time will all have regular characteristics due to the theme change. Therefore, in general, the theme feature extraction technology for the forum is not only to process the text, the above elements combine and extract their characteristics, with thematic characteristics of the data and other data separate.

There are a number of options for clustering [14], namely concept derivation, group derivation words, group word repetitions, group related concepts, the co-occurrence rules are group co-occurring concepts. In this paper, we choose co-word phrase clustering based on k-means algorithm.

4) *Emotion*

Emotions are the attitude and experience that people make about whether or not objective things satisfy their needs [15]. Wilson et al. [16] believe that the task of emotional analysis is to identify positive and negative views and emotions. To put it in a nutshell, sentiment analysis is to classify a passage of text, sentences or words into categories of positive, negative, happy, angry, sad, happy, good, fearful, evil, and scared. The emotion studied in this paper is based on the posts, the students' postings vary in length, but the rest of the postings except the topic posts are generally short, so the affective analysis involves two types of sentences and chapters, but mainly the sentence-level sentiment analysis [17].

5) *Social network analysis*

Social Network Analysis (NA) takes the social actors and their mutual relations as the research contents, and describes the mode of the actors' relational modes and their implications for the actors and the entire group [18]. Social network analysis was originally a sociological research approach to study the relationship, the content and intensity of the main link, and thus form the relevant subgroup network.

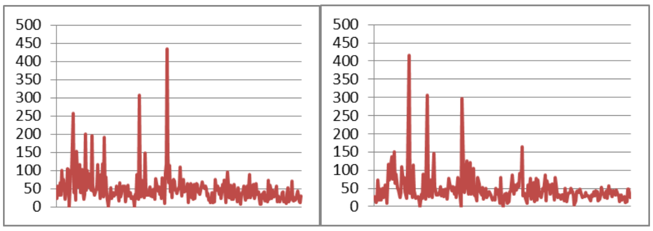
IV. Empirical Analysis and Practical Reflection

Based on the interactive discourse analysis mode constructed in this paper, this paper takes a college graduate's educational technology research method as an example, and draws two groups of forum dialogue posts in this course, and analyzes them one by one. The duration of the course is 18 weeks. There are 8 groups in the class. Each group has conducted 7 topic postings in the forum. This study draws the topic posts of two groups for research and analysis. The forum part of the course is based on the progress of the teacher's research, and each group will post and exchange topics according to their own research topics. The selected subjects in Group I and II respectively posted a total of 293 articles, and a total of 226 posts were posted under the theme of Group B, and the landlord was a teacher, which was not included in the posting statistics and research scope.

Before the data analysis, the data cleaning work needs to be completed. Firstly, the obtained data is processed by hidden personal information, and the student number is replaced by other numbers that do not represent personal information. In this study, it is replaced by s1, s2. Wait. Secondly, the information such as duplicates and false transmissions in the data is deleted. In the case of sentiment analysis and topic mining of texts, the literature review, research plans, etc. submitted by each group are not included, because this paper focuses on the analysis of the educational information generated during the dialogue process. The length will affect the analysis results.

A. *Statistical Description*

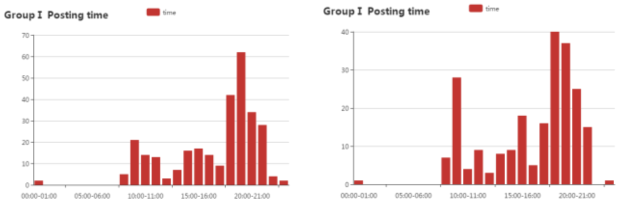
The statistical description describes the time of posting, the number of posts, and the number of words in the post. In one semester of study tasks, members of Group I posted a total of 260 posts, 43.3 per capita, and 198 members of Group II posted 39.6 posts per person. The topic posts in the two groups are 293 and 226 respectively. These all indicate that the online forums according to Group I are more active than Group II. Except for the group topic posts, the number of words in the other posts is mostly between 0 and 100. In comparison, there are more than 150 words in the I group, as shown in Fig. 5.



a) Group I b) Group Ⅱ

Fig. 5. Posting length statistics.

According to statistics, the posting date of the two groups is generally concentrated on the day of the experimental course or the day after, and the posting time is concentrated between 6 pm and 10 pm. It was also found that members of Group II posted a higher amount of postings between 9:00 and 10:00 in the morning, which prevented the completion of the course tasks on time from the side reaction group II. In comparison, members of Group I are more willing to "overtime" to complete the course tasks on the same day, because in the time statistics, it is found that the frequency of postings from Group 10 in the evening to 1 in the morning is higher than in Group II. See Fig. 6 for details.



a) Group I b) Group Ⅱ

Fig. 6. Group I experiment post time statistics.

Through the above data analysis, it is found that the student's forum learning behavior mostly occurs at the beginning or end of the day's study. Because the posting length is also short, the use of piecemeal time learning is more in line with the student's rhythm. In addition, the topic posts of Group I are more likely to cause discussion. The specific reasons need further analysis. It may be that the members of the group are active, and it may be that the learning content of the group is more interesting for discussion.

B. *Emotional Analysis*

The emotional state can reflect a student's attitude towards the learning task. Through the emotional analysis of the student posting, they find that their speeches in the forum are generally positive. The emotional differences of the topic posts of different groups are not large, and they all remain at 0.7-0.8. between. However, there are still some differences between the posts and the emotions of different team members. The specific results are shown in Fig. 7:

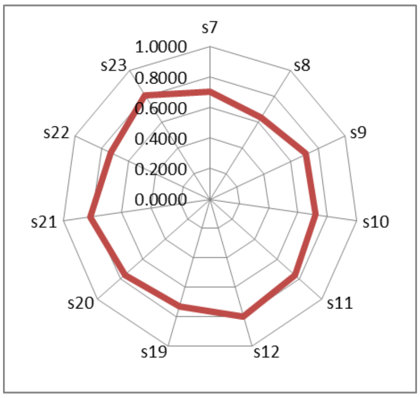


Fig. 7. Group member sentiment analysis.

It is easy to analyze from the radar chart, the overall sentiment of students posting is positive, but the emotional value of individual students is low, such as S8. In this case, the teacher can guide the students to think about the research questions from the more familiar content. In the process of answering questions and teaching, they should also pay attention to the research topics that students are more interested in.

C. Social Network Analysis

The tool for social network analysis of the interaction of forum posts is Ucinet, and the results are shown in Fig.8. The red equilateral triangle represents group I, the white inverted triangle represents group II, and the gray square represents other members. It can be clearly seen from the figure that there are always some star nodes, some edge nodes and some common active nodes in the interaction of each group, which is an important means for teachers to understand student activity.

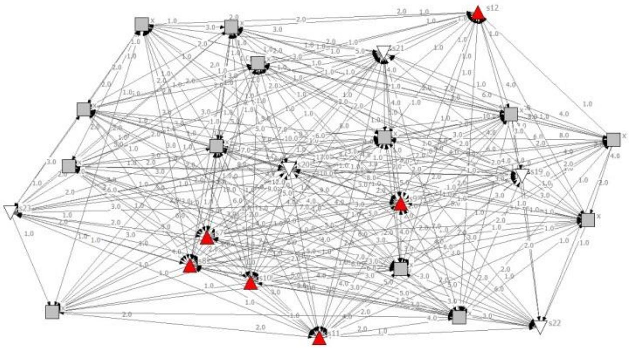
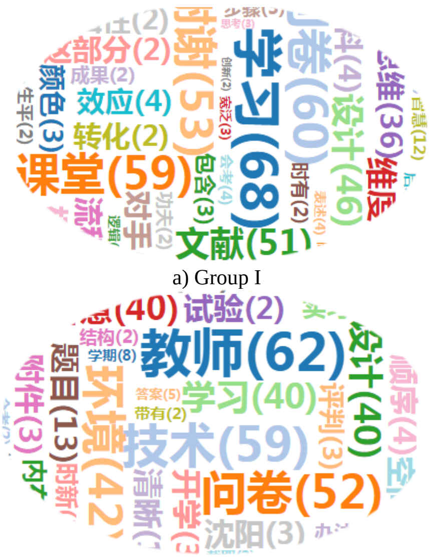


Fig. 8. Forum student interaction.

Researchers at Montesquieu University and the University of Edinburgh found that learners' learning outcomes are more pronounced when learners express narrative style discourses with a refined, simple grammatical structure and abstract vocabulary. The central position is more prominent. However, in this study, the central students did not achieve significant learning outcomes. By further studying its language expression, although the student has a high frequency of speeches and a large number of posts, the question and answer “quality” needs to be improved. It may be that the learning ability needs to be improved. At this time, the teacher can guide the student to think deeply.

D. Theme Mining

Use the NLPIR integration tool to extract the keywords of the forum stickers, as shown in Fig. 9. NLPIR visually describes the word frequency and its co-occurrence in the form of a graph, which helps teachers and students to have an intuitive understanding of the forum topic.



a) Group I

b) Group Ⅱ

Fig. 9. Keyword analysis.

In the specific use, the keyword information can be used to check whether the discussion sticker deviates from the theme. If you don't know the subject, you can look for the aggregated topic of the discussion post. It can be seen that the theme of Group I is the subject of the class, and the group of II is the subject of teachers and technology. In comparison, Group I students received more clear suggestions, which can be inferred from the high-frequency vocabulary of “thank you,” and Group II students should have encountered obstacles in conceptual issues because “documents” are “clear” and so on. The word is high frequency, and the word "design" is also used as a high-frequency vocabulary. It can be inferred that there is a problem in the design aspect of the research. At this point, the teacher should focus on the program of the group II students, and analyze and answer the problems that arise.

E. Text Meaning

The mining of the meaning of the text is done by manually measuring the quality of the forum posts. The same topic posts of the two groups are analyzed according to the dimension of the dialogue content determined in advance. The results are shown in Table I:

TABLE Ⅰ: FORUM STATISTICS

Num	Category	Number (Unit / Article)			
		group Ⅰ		group Ⅱ	
1	Share and clarify	63	21. 50%	60	26. 55%
2	Meaning negotiation	112	38. 23%	73	32. 30%
3	Test correction	82	27. 99%	71	31. 42%
4	Achieve and application	36	12. 29%	22	9. 73%
5	Cognitive conflict	0	0. 00%	0	0. 00%
6	Total	293	100%	226	100%

It can be seen from the table that category 1 and category 2 got the largest number of posts, which shows that in the forum dialogue each group is more likely to ask questions about the contents of other groups while others clarify the meaning of the questions. For deeper inspection and trimming, applications and even cognitive conflicts are rare.

Therefore, it is inferred that the students' thinking on the theme of the forum is rather superficial. The topic of the forum cannot attract the attention and thinking of the students very much. Combining with their own understanding of the status of the curriculum, this is related to the limited time for the participants to devote more energy to their activities.

V. Conclusion

From the empirical results, the mode's indicators can fully and accurately reflect the student's forum learning behavior, this mode still needs to be constantly tested and revised, with the development of technology can also have a more complete index system. And different types and levels of disciplines should have their own characteristics, and the topics and forms of effective interactive dialogue should be different. Therefore, based on the establishment of a standard semantic library, a new semantic library with subject characteristics should be added for further analysis. This article does not make the technical integration and scale application of the proposed mode, which is the biggest deficiency of this study, and will continue to be improved and amended next.

References

[1] Y. Jung and J. Lee, "Learning engagement and persistence in massive open online courses (MOOCS)," *Computers & Education*, vol. 122, pp. 9-22, Jul 2018.

[2] E. Saitta, "Virtual practice in STEM teaching assistant professional development: Using a mixed-reality teaching simulator to enhance classroom discourse in student-centered learning environments," *Abstracts of Papers of the American Chemical Society*, vol. 255, Mar 18 2018.

[3] J. Griswold, L. Shaw, and M. Munn, "Socratic seminar with data: A strategy to support student discourse and understanding," *American Biology Teacher*, vol. 79, pp. 492-495, Aug 2017.

[4] B. Oyarzun, J. Stefaniak, L. Bol, and G. R. Morrison, "Effects of learner-to-learner interactions on social presence, achievement and satisfaction," *Journal of Computing in Higher Education*, vol. 30, pp. 154-75, April 2018.

[5] H. Shan, "Lifelong education and lifelong learning with Chinese characteristics: A critical policy discourse analysis," *Asia Pacific Education Review*, vol. 18, pp. 189-201, 2017.

[6] R. Kling and C. Courtright, "Group behavior and learning in electronic forums: a sociotechnical approach," *Information Society*, vol. 19, pp. 221-35, July-Aug. 2003.

[7] R. Martinez-Maldonado, Y. Dimitriadis, A. Martinez-Mones, J. Kay, and K. Yacef, "Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop," *International Journal of Computer-Supported Collaborative Learning*, vol. 8, pp. 455-85, Dec. 2013.

[8] X. She, P. Jian, P. Zhang, and H. Huang, "Leveraging hierarchical deep semantics to classify implicit discourse relations via a mutual learning method," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, May 2018.

[9] Z. Yusoff, A. Kamsin, S. Shamshirband, and A. T. Chronopoulos, "A survey of educational games as interaction design tools for affective learning: thematic analysis taxonomy," *Education and Information Technologies*, vol. 23, pp. 393-418, Jan. 2018.

[10] R. Moreno, "Web forum retrieval and text analytics: A survey," *Foundations and Trends in Information Retrieval,* vol. 12, pp. 2-5, 2018.

[11] Y. Adachi, *Visual Language Design System Based on Context-Sensitive NCE Graph Grammar*, Chichester, U.K.: Wiley, 2004, pp. 45-47.

[12] C. Bodong, C. Yu-Hui, O. Fan, and Z. Wanying, "Fostering student engagement in online discussion through social learning analytics," *Internet and Higher Education*, vol. 37, pp. 21-30, April 2018.

[13] D. Chen, Y. Dai, and Z. Wang, "Survey of research on forum topic mining," *Computer Engineering and Applications*, vol. 53, pp. 36-44, Aug. 2017.

[14] G. Li and Y. Long, "A study on hotspots and trends of international technology foresight in recent ten years (2004-2013)," *J. Knowledge of Library and Information Service*, vol. 3, pp. 104-116, 2014.

[15] C. Lin, Z. Yang, and X. Huang, *Psychology Speech*, Shanghai: Shanghai Education Press, 2003, pp. 100-115.

[16] T. Wilson and J. Wiebe, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver: ACL, 2005, pp. 347-354.

[17] X. Tang and G. Liu, "A review of fine grained emotion analysis," *Journal of Library and Information Service*, vol. 61, pp. 132-140, 2017.

[18] S. Wasserman and K. Faus, *Social Network Analysis: Methods and Applications*, Cambridge, U.K: Cambridge Press, 1994, pp. 35-40.

Zhifeng Wang received the BEng degree in electronic engineering from China University of Geosciences in 2008, and the PhD degree in electronic engineering from South China University of Technology, China, in 2013. He was a joint training PhD student in the Computer Science Department of Carnegie Mellon University, Pittsburg, PA, during 2010 to 2011. He is now an associate professor in the School of Educational Information Technology of Central China Normal University. His research interests include learning analytics, data mining, machine learning, and learning theory.

Rong Zhao received the BEng degree from Central China Normal University in 2016. She is currently a postgraduate in the School of Educational Information Technology. Her main research areas are learning analysis models and practical applications.