International Journal of Knowledge and Language Processing
Volume 12, Number 1, **2021**

www.ijklp.org
KLP International ⓒ2021 ISSN 2191-2734
pp.11–21

The Research Focus and Trends in Educational Text Mining：Quantitative Analysis of Literature

Meng Wang, Qianqian Zhang, Yaru Fu, Zhijia Mou

Department of Educational Technology, Jiangnan University
No. 1800, Lihu Road, Binhu Distict, Wuxi, Jiangsu, China
wangmengly@163.com

Received August 2020; revised August 2020

ABSTRACT. *Educational text mining is an important research field in learning analytics. The analysis of the texts generated in the learning process helps to better master the students' inner psychological characteristics such as emotion, thinking and cognition, and provide evidence support for precise teaching design. In this study, we analyzed 111 articles on educational text mining in Web of Science core database and Elsevier ScienceDirect from 2010 to 2019 using visual softwires such as CiteSpace5.6, Bicomb2.0, HistCite and SPSS25.0, and made a quantitative analysis from the aspects of citation relationship, cooperation relationship and time sequence relationship. As an overview, this paper intends to describe what is the latest development, how they have evolved, and what are the research trends in educational text mining.*

Keywords: Educational text mining; Natural language processing; Visual analysis

1. **Introduction.** In recent years, with the development of educational big data and artificial intelligence, learning analytics has gained wide attention of scholars. In the early stage of learning analytics, scholars mainly focused on learning behavior data, but few on other types of data. In digital learning, not only the interactive behavior data, but also the interactive discussion text data are produced in this process. Thus, the analysis of text data is very helpful to discover the internal psychological characteristics of learners. As an important research field in data mining, text mining is a process of extracting the available information and knowledge from a large number of texts. With the diversified development

of learning data and the in-depth application of text mining, the analysis based on learning text data has been widely concerned by scholars, and lots of theoretical and practical explorations have been carried out. Thus, it promotes the emergence of educational text mining. Educational text mining is a process of extracting effective information and knowledge from various interaction, evaluation and reflection text data generated in the process of teaching and learning activities.

The development of natural language processing provides a new technical method for educational text mining. To clarify the research status of educational text mining and find out its future research orientation is of great practical significance for learning analytics. In this paper, through the analysis of the literature on text mining in recent ten years, the main research topics in this field are extracted to provide ideas and reference for the follow-up research.

2. Data Collection and Methodology.

2l **Data Collection**. We retrieve documents from Web of Science core database and Elsevier ScienceDirect（2010-2019）by searching the titles on keywords："educational text mining", "text analysis & online course", "text analysis & educational online discussion, discourse analysis & online learning" and "content analysis & distance education". By eliminating the duplicate documents and non-research documents such as conference reports or news, we finally keep 111 documents for this study.

22 **Methodology**. In this study, both quantitative and qualitative methods are employed in the analysis of social network, clustering of the literature and visualization of knowledge graph. We use 4 tools to manage the documents and do statistical analysis which includes CiteSpace5.6, BICOMB2.0, Histcite and SPSS25.0. The research procedures are as follows: (1) Retrieve the documents from the Web of Science and Elsevier ScienceDirect database, and the basic information such as title, author, keywords, abstract and journal name are downloaded and imported into the document management software for unified management; (2) Histcite is used to analyze the citation relationship of literature to find out the core literature in educational text mining; (3) CiteSpace5.6 is used to analyze the author's cooperation relationship and find out the core authors and research groups; (4) The keywords in the literature are extracted, and the co-occurrence analysis of keywords is conducted to find out the high-frequency keywords; (5) Clustering analysis is used to analyze the research topics in educational text mining and the relationship between them; (6) Based on the analysis of time series graph, the research hotspots of educational text mining are divided by year, and the future research trends are predicted.

3. Experiments and Analysis.

3l **Keywords Co-occurrence**. By analyzing the keywords co-occurrence, scholars can quickly find out the core research hotspots and research characteristics in this field. Using Citespace5.6, we generate the keywords co-occurrence map including 314 nodes and 759 lines. The keywords such as "text mining", "educational text mining", "content analysis"

are deleted, and the threshold value is set to 2. The keywords co-occurrence map is shown in Figure 1. The size of the font indicates the frequency of keywords: the larger the font, the higher the frequency of keywords. The connection between nodes indicates the centrality of keywords: the more lines, the higher the centrality of keywords.

It can be seen the keyword with the highest co-occurrence frequency is "sentiment analysis", with a total of 10 times; the second is "E-Learning", with the frequency of 9. The frequencies of "learning analysis", "online learning", "machine learning", "online discussion", "natural language processing" are all greater than 6. In terms of keywords connection, "E-Learning", "machine learning" and "sentient analysis" are most closely connected. And it shows that machine learning is mainly used to analyze students' sentiments in digital learning environment. Secondly, "online discussion" and "Computer Supported Collaborative Learning" are closely connected, and it indicates that computer supported collaborative learning and its text generation are the main research topics. In addition, keywords such as "qualitative research" and "research method" show that educational text mining is often used to find out the research hotspots and research history of a certain field.

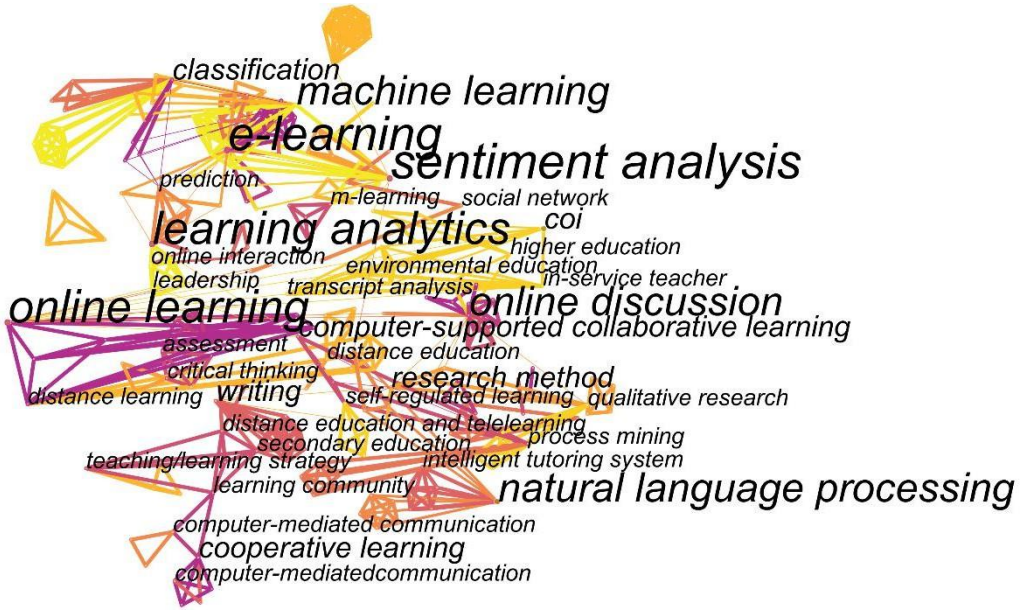

FIGURE 1. KEYWORDS CO-OCCURRENCE ANALYSIS

32 **Keywords Clustering.** For the keywords clustering, we imported the keywords co-occurrence matrix into SPSS25.0 and select "system clustering method" under "classification" option. The clustering is set as "inter group clustering", and the measurement interval is "Euclidean distance". Then the keywords clustering map of educational text mining is generated, as shown in Figure 2.

The keywords are listed in the left side. The numbers in vertical axis represent the sequence number of the variables in the whole set, and the numbers in horizontal axis represent the distance between variables. Keywords are mainly clustered into five categories. Cluster 1 is the literature source and analysis methods in educational text mining.

Cluster 2 is the analysis of students' behavior, including the behavior patterns in online learning and the interaction in online discussion. Cluster 3 is the presentation methods in educational text mining, which currently includes user portraits and models, concept maps [1]. Cluster 4 is the analysis and evaluation of learning outcomes, such as predicting the dropout rate of students' online learning based on the students' discussion texts and evaluating students' writing level based on natural language processing technology. Cluster 5 is students' cognition and emotion analysis, such as students' knowledge construction process and emotion changes.

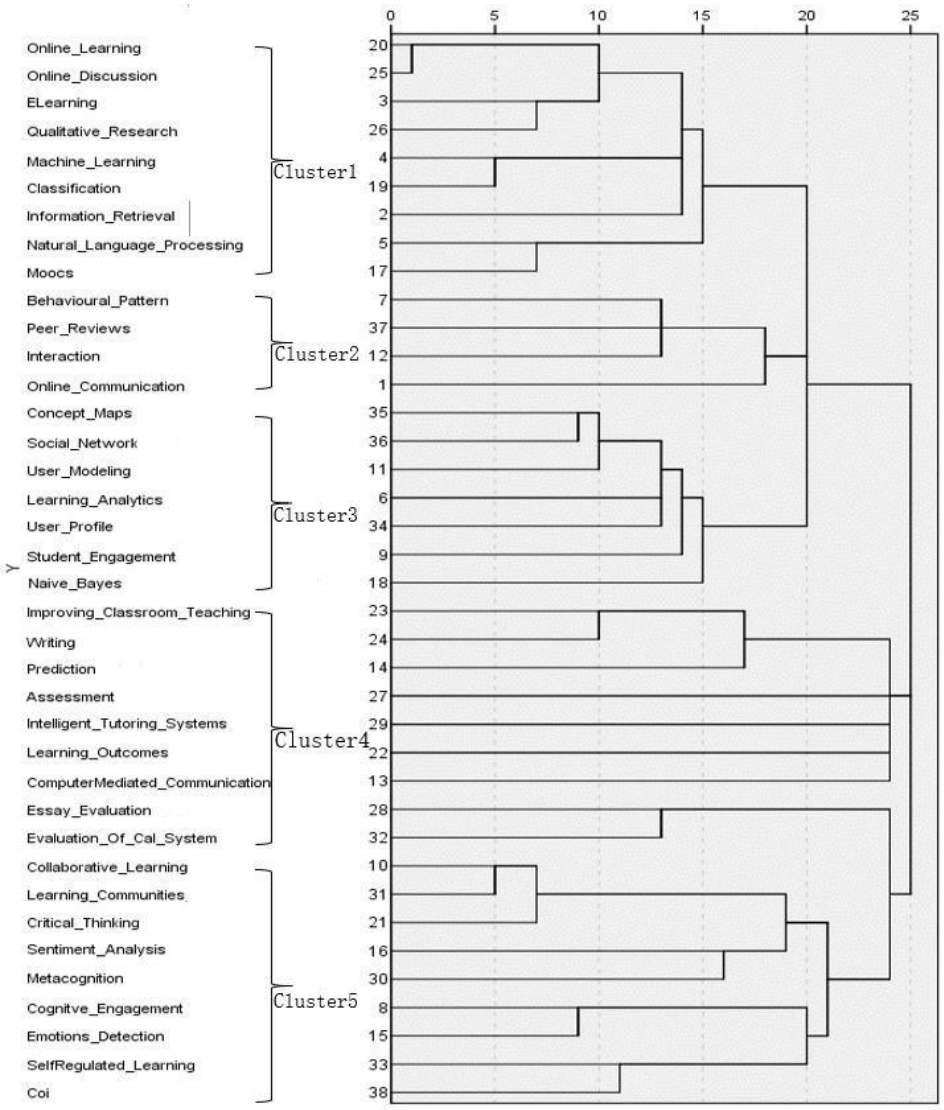

FIGURE 2. KEYWORDS CLUSTERING ANALYSIS

33 **Highly Cited Literature**. Since the quantity of the literature in educational text mining and the time span are both large, it is of great significance to clarify the mutual citation relationship and find out the key literature. In this section, we use Histcite to analyze the

citation relationship in the literature. On the parameter setting, the display node is set as "GCS" (total citation), and the display quantity is set as 70. The results are shown in Figure 3. The vertical numbers represent the publication year, the circles represent the literature, and the number in the circle represents the sequence number of the literature in the whole set.

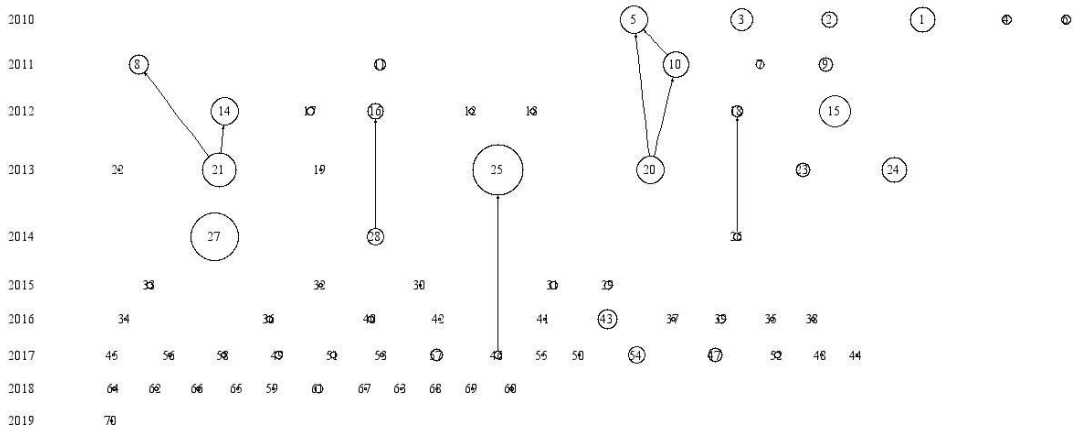

FIGURE 3. HIGHLY CITED LITERATURE

Note that there is less citation relationship between the literature, which may be due to the fact that the current research on educational text mining involves many fields and the relationship between them is not so strong. The most frequently cited literature is No. 25, with a total of 127 citations, which is the "Predicting students' final performance from participation in on-line discussion" written by Cristóbal Romero in 2013. This paper presented how to obtain a more powerful and interpretable model through centroid and class association rules. The research shows that the features such as the number of messages sent, the number of words in the evaluation text, the average evaluation value of the text, the centrality and degree of the text are the most important to predict the final performance of students [2]. The second frequently cited literature is No. 27, with a total of 113 citations, which is "Sentiment analysis in Facebook and its application to e-learning " written by Alvaro Ortigosa in 2014. The authors of this paper put forward a new method of emotion analysis on Facebook, and developed a program named Sentbuk to retrieve the emotional polarity information (positive, neutral or negative) in the user's published texts to detect significant sentiment changes. It is pointed out this method can be applied to the field of online learning by supporting personalized learning and recommending learning material based on learners' emotional state [3]. Except for the above two papers, the rest ten highly cited literature are listed in Table 1.

TABLE 1. TEN HIGHLY CITED LITERATURE

No.	Title	Author	Publication date	Cited frequency
21	Examining students' online interaction in a live video streaming environment using data mining and text mining	Wu He	2013	59
15	Examining mobile learning trends 2003-2008: a categorical meta-trend	Jui-Long Hung & Ke	2012	51

	analysis using text mining techniques	Zhang		
5	Exploring the behavioural patterns in project-based learning with online discussion: quantitative content analysis and progressive sequential analysis	Huei-Tse Hou	2010	41
14	Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics	Jui-Long Hung	2012	41
20	Analyzing knowledge dimensions and cognitive process of a project-based online discussion instructional activity using Facebook in an adult and continuing education course	Peng-Chun Lin etc.	2013	41
10	A case study of online instructional collaborative discussion activities for problem-solving using situated scenarios: An examination of content and behavior cluster analysis	Hou Huei-Tse	2011	35
1	Understanding telecollaboration through an analysis of intercultural discourse	Meei-Ling Liaw & Susan Master	2010	33
24	Natural language processing in an intelligent writing strategy tutoring system	Danielle S. McNamara etc.	2013	33
3	A quantitative multimodal discourse analysis of teaching and learning in a web-conferencing environment - The efficacy of student-centred learning designs	Matt Bowera & John G.Hedberg	2010	26
8	Using text mining to uncover students' technology-related problems in live video streaming	M'hammed Abdous & Wu He	2011	19

3.4 **Author's Cooperation Relationship.** By analyzing the author's cooperation relationship, we can find out the leading scholars in the field of educational text mining and their cooperation to know better the contributions of different authors. First we import the data into Citespace5.6, and set the following parameters: the time range as "2010-2019", the time slice as "one year", the node type as "author", and the "threshold" in "Selection Criteria" as (1, 1, 20). Then we get the author cooperation graph with 281 nodes and 409 lines, as shown in Figure 4. As can be seen, Huei-tse Hou has the highest frequency (3 times), but there are few connections around it, which shows that it has less cooperation with other scholars. The second are Cristóbal Romero, Wu He, Jui-long Huang, Kui Xie, Maija Aksela and Ming Ming Chiu, with the frequencies of 2 times. Except for the close relationship between Cristóbal Romero and other scholars, the other scholars with higher frequency have less cooperation. This is more consistent with the current research status of educational text mining. Scholars use different methods such as natural language processing, discourse analysis, content analysis to analyze the texts of different disciplines or platforms. Thus, the application of educational text mining is more extensive and diversified, and the cooperation between different scholars is relatively loose. Moreover, in

the existing mature cooperation relationship, the research directions are also very different from each other. For example, the group led by Huei-tse Hou focuses on the behavior patterns and limitations of students in online forums. The group led by Cristóbal Romero mainly focuses on predicting students' performance in online learning, while the group led by Wu He focuses on the quality of communication and interaction among students in the process of online learning.

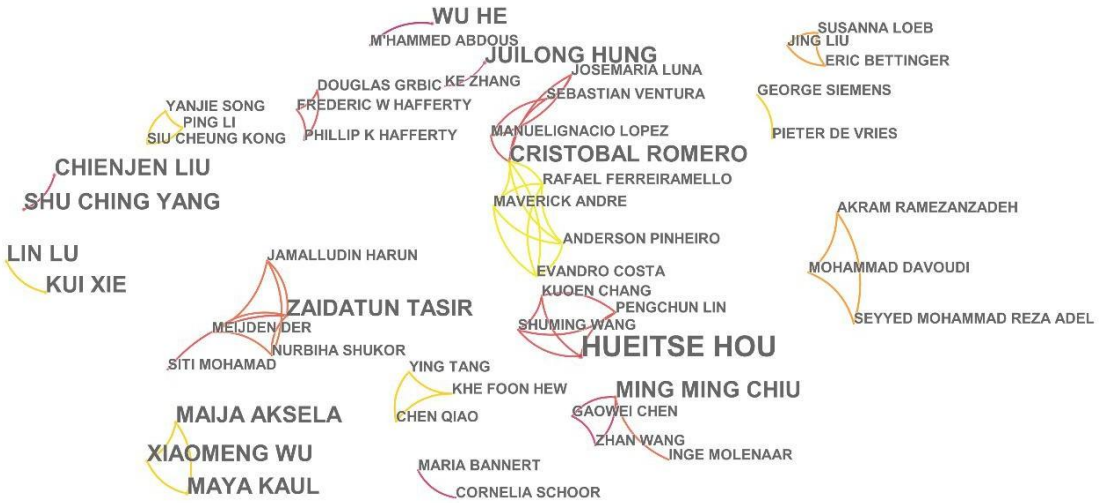

FIGURE 4 AUTHOR'S COOPERATION RELATIONSHIP

35 **Evolution Trends Based on Time Series Graph.** Using CiteSpace 5.6, we set the "node types" as "keyword", and the other parameters remain unchanged. Finally, a graph containing 314 nodes and 754 lines is generated. Then we select the visualization type as "timezone view". Note that only keywords with frequency greater than 2 will be displayed in order to visualize the keywords evolution trends clearly, as shown in Figure 5.

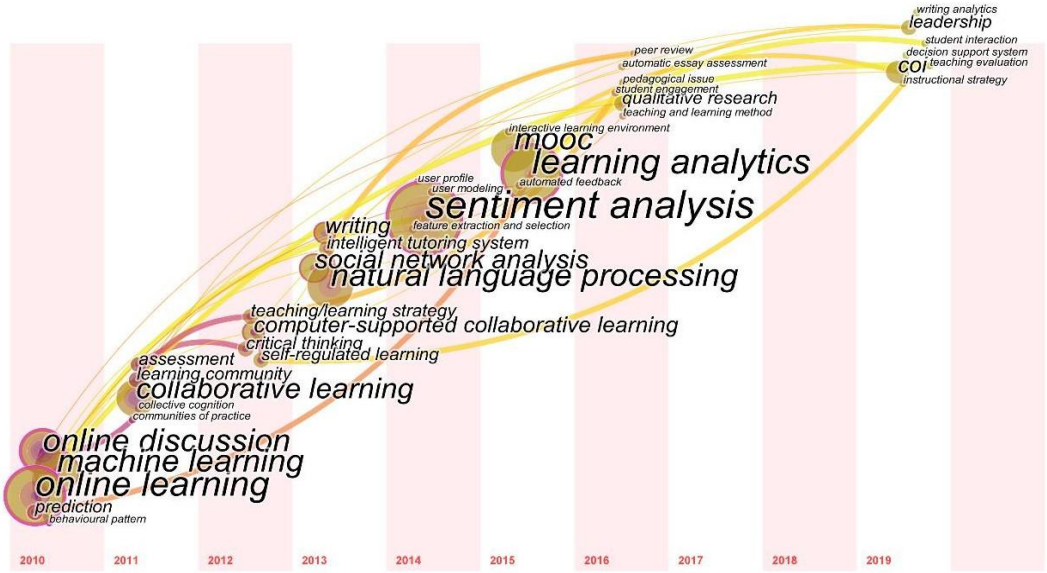

FIGURE 5 EVOLUTION TREND BASED ON TIME SERIES GRAPH

We can see that early research on educational text mining focuses on collaborative learning and online discussion. With the development of text mining technology, sentiment analysis and curriculum teaching have become the mainstream of mid-term research, and now it focuses on knowledge construction and cognitive level. In the early stage, scholars explored collaborative learning and interactive discussion. For example, Matt Bower et al. proposed a quantitative multimodal discourse analysis method to analyze students' online collaborative learning to know better the influence of task theme, activity design and interface selection on students' interaction and collaboration in the web-conference environment [4]. Taking Taiwan and the United States' pre-service normal education students as an example, Meei-Ling Liaw studied the discourse characteristics of learners in the forum interaction, as well as the interaction mode and type between cross-cultural interlocutors, to explore how to carry out collaborative and cross-cultural learning in the process of distance learning [5]. Huei-Tse Hou selected 23 college students as a sample to analyze how to carry out discussion activities under the case scenarios and problem-solving tasks given by teachers, so as to test the effectiveness of case-based online collaborative discussion in higher education courses [6].

In sum, research in educational text mining shows the following three orientations: (1) Data integrity: from focusing on part of the text to the overall text in the learning process; (2) High-level analysis: from the text-based knowledge analysis to the text-based learning psychology and thinking characteristics analysis; (3) Intelligent algorithm: from simple text word frequency statistics to machine learning based natural language processing technology. Moreover, the research on educational text mining develop from single channel discourse analysis to multi-channel context analysis, and systemic functional linguistics is used to deconstruct the meaning generated in the interaction text.

4. **Analysis of Research Hotspots in Educational Text Mining**. Based on the different application scenarios of educational text mining and qualitative analysis of research topics, we found that the current research on educational text mining focuses on the following five aspects.

(1) Analysis of learning behavior and cognition based on curriculum texts

The curriculum texts include texts in forum, comment, dialogue, assignment and so on. As an output generated by students in the learning process, curriculum texts can reflect the learning and thinking process of students and help the teachers understand the way how the students deal with information. By analyzing students' online curriculum texts, students' learning behavior and cognitive process can be clearly presented. Based on curriculum texts, activity texts, evaluation texts, and the number of click and visit in online learning, students' learning behavior can be evaluated and predicted. For example, Rianne Conijn et al. took the MOOC based hybrid curriculum as an example, and used the students' activity texts, activity frequency and activity sequence to predict the performance of students. The research has shown that online activities can be used as an important indicator to evaluate the students' performance in mixed courses [7].

(2) Sentiment analysis based on interaction texts in classroom

Learners' sentiment changes directly affect their attitudes towards the course and the final learning outcomes. Sentiment analysis based on classroom interaction texts mainly uses data mining techniques such as classification, clustering and neural network to retrieve the sentiment keywords in the texts. The results can be used as an important indicator to evaluate students' learning process as well as an important basis for improving teaching. At present, the research on sentiment analysis mainly focuses on two aspects. First, by analyzing the reflective texts or online discussion texts generated by learners in online learning, the learners' sentiments can be detected, and targeted intervention can be carried out as well. For example, Kamisli Ozturk et al. analyzed the interaction texts and the Twitter texts published by students participated in the course, and used Naive Bayes to classify the sentiments into three categories: positive, negative and neutral, as so to test the acceptance of distance education [8]. Ling Wang et al. proposed a semantic analysis model to track MOOC learners' sentiment from real-time data such as assignments and reviews in order to analyze students' acceptance of courses and predict the course completion rate in different learning stages in real time. And it is also a solution for personalized teaching of MOOC [9]. Second, scholars tend to use machine learning or text mining techniques to develop sentiment analysis tools to evaluate students' emotion in learning. For example, Samuel Nelson et al. used LDA to explore students' opinions on various aspects of the course, and the results were visualized and fed back to teachers [10]. Feng Tian et al. proposed an emotion recognition framework based on interaction texts, and used Random Forest algorithm to classify learners' emotions [11].

(3) Analysis of knowledge construction based on collaborative text

Online collaborative knowledge construction is an important topic of the socialization of e-learning. However, the current situation in collaborative knowledge construction is "No cooperation in activities, and no construction or low level of construction". Therefore, how to analyze the process of students' knowledge construction and evaluate their knowledge construction levels has become a critical issue. It is the main form of online collaborative knowledge construction to analyze the texts, the number of posts, the length of discussion and other data submitted by students in collaborative learning to evaluate students' knowledge construction levels based on the existing knowledge construction framework. For example, Liu Chien-Jen et al. investigated the levels of knowledge construction in asynchronous discussion of online courses and analyzed students' attitude and perception of online discourse based on the "inquiry learning community" model proposed by Garrison [12]. Fatemeh Nami et al. explored the knowledge construction process of teachers in asynchronous communication discussion. Based on the existing classification framework, the cognitive changes of teachers in cognitive existence, social existence and teaching existence were identified [13].

(4) Automated writing evaluation based on written text

Writing evaluation is an indispensable part in language teaching, and lots of work has been done by scholars. However, the evaluation of writing has always been in the dilemma of time-consuming, laborious and low reliability [14]. The development of natural language processing and text mining technologies provide new solutions for writing evaluation. The

statistical methods, such as word frequency, lexical diversity, syntactic similarity, syntactic complexity, latent semantic analysis and semantic coherence analysis, can automatically evaluate students' written texts and detect plagiarism. Therefore, scholars develop or utilize the existing platforms to analyze students' written texts and give them appropriate feedback so as to improve their writing ability. For example, Danielle McNamara et al. extended the current writing evaluation model and proposed a five-dimension framework to evaluate writing ability, namely vocabulary, syntax, cohesion, rhetoric and readability. Based on the existing intelligent tutor system "the writing pal" and the evaluation framework, they analyzed the students' written texts and provided valuable feedback in the writing tutor system [15]. Gokhan Akcapinar et al. reduced plagiarism in online homework by providing automatic feedback using text mining analysis [16]. Soobin Yim et al. analyzed the written texts of middle school students in the second language environment to verify the students' collaborative process [17].

5. **Conclusion**. As an important research field in learning analytics, educational text mining can provide convincing evidence support for the data analysis of the whole learning process. This study made a quantitative analysis from the aspects of citation relationship, cooperation relationship and time sequence relationship using co-occurrence analysis and clustering analysis on the literature in educational text mining. It can be seen the current research mainly focuses on text analysis and platform development which explores the behavior, cognition, emotion and ability of students in online courses. For the data source, it is necessary to further combine other learning behavior data for comprehensive analysis. For the analysis methods, it relies on natural language processing and deep learning technology for semantic computing and understanding. For the analysis objects, it expands from learning texts to interaction texts between teachers and students. Automated processing of massive learning texts can improve the efficiency of text processing and the depth of text mining, and provide strong support for learning and teaching.

Acknowledgment. This work is partially supported by Chinese Ministry of Education Research Projects of Humanities and Social Sciences (No. 18JYC006). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Ferreira-Mello R, André M, Pinheiro A, et al. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6): 1332-1341, 2019.

[2] Romero C, López M I, Luna J M, et al. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68(10): 458-472, 2013.

[3] Ortigosa A, Martín J M, Carro R M. Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, 31(2): 527-541, 2014.

[4] Liaw M L, Bunn-Le Master S. Understanding telecollaboration through an analysis of intercultural discourse. *Computer Assisted Language Learning*, 23(1): 21-40, 2010.

[5] Hou H T. A case study of online instructional collaborative discussion activities for problem-solving using situated scenarios: An examination of content and behavior cluster analysis. *Computers & Education*, 56(3): 712-719, 2011.

[6] Esparza G G, Díaz A P, Canul-Reich J, et al. Proposal of a Sentiment Analysis Model in Tweets for improvement of the teaching-learning process in the classroom using a corpus of subjectivity. International Journal of Combinatorial Optimization Problem s and Informatics, 7(2): 22-34, 2016.

[7] Conijn R, Van den Beemt A, Cuijpers P. Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, 34(5): 615-628, 2018.

[8] Ozturk Z K, Cicek Z I E, Ergul Z. Sentiment Analysis: An Application to Anadolu University. *Acta Physica Polonica*, 132(3): 753-755, 2017.

[9] Wang L, Hu G, Zhou T. Semantic analysis of learners' emotional tendencies on online MOOC education. *Sustainability*, 10(6): 1921, 2018.

[10] Cunningham-Nelson S, Baktashmotlagh M, Boles W. Visualizing Student Opinion Through Text Analysis. *IEEE Transactions on Education*, 62(4): 305-311, 2019.

[11] Tian F, Gao P, Li L, et al. Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems. *Knowledge-Based Systems*, 55(1): 148-164, 2014.

[12] Liu C J, Yang S C. Using the community of inquiry model to investigate students' knowledge construction in asynchronous online discussions. *Journal of Educational Computing Research*, 51(3): 327-354, 2014.

[13] Nami F, Marandi S S, Sotoudehnama E. Interaction in a discussion list: An exploration of cognitive, social, and teaching presence in teachers' online collaborations. *European Association for Computer Assisted Language Learning*, 30(3): 375-398, 2018.

[14] Landauer T K. Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, 10(3): 295-308, 2003.

[15] McNamara D S, Crossley S A, Roscoe R. Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45(2): 499-515, 2013.

[16] Akçapınar G. How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers & Education*, 87(9): 123-130, 2015.

[17] Yim S, Warschauer M. Web-based collaborative writing in L2 contexts: Methodological insights from text mining. *Language Learning & Technology*, 21(1): 146-165, 2017.