Application Analysis of Data Mining Technology in the Field of Learning Analysis in China—From the Perspective of Content Analysis

Shurui Gao[*], Luyue Li, Juan Wu

School of Educational Technology, Beijing Normal University, Beijing, China.

Abstract

Learning analysis has gradually become the focus of attention of the majority of educators, and data mining has brought more possibilities for learning analysis. This study takes CNKI literature as the research object, searches "data mining" and "learning analysis" as keywords, and the content analysis method is used to explore the application status of data mining technology in learning analysis researches. This study found that relevant research scenarios focus on online platforms and higher education. The research topics are mainly theoretical exploration and model construction to explore the rules of teaching and learning and analyze learners. Data collection gradually diversified, focusing on process data collection. Analysis method tends to be the comprehensive application of professional data mining method. In the future, the scope of research objects can be further expanded, classroom performance data, unstructured data and physiological data can be collected for analysis, and consider data privacy to guarantee the legal rights and interests of the participants.

Keywords

Data Mining, Learning Analysis, Content Analysis

1. Introduction

In recent years, learning analysis has become the focus of attention of a vast majority of educators. Learning analysis refers to the measurement, collection, analysis and report of data related to students' learning behavior and learning environment, and is used to understand and optimize learning and the environment in which it occurs. It can be seen that data is an important premise of learning analysis (Lin & Yang, 2017), and data analysis is an even important link of learning analysis research. As a typical technology of data analysis, data mining technology adds a new possibility for learning analysis researches. The concept of data mining appeared in the late 1980s, it is an advanced treatment process that extracts potential, effective and readable patterns from huge amounts of data according to the established goals (Zhong, Ma, Zhang, & Yu, 2001), as a multidisciplinary integrated product, it use artificial intelligence, machine learning, pattern recognition, statistics, database and visualization technology to automatically analyze data and gain potential implicit knowledge (Wang, 2013). With the increasing number of learning data acquisition channels and the increasing amount of data generated by learners, data mining technology has gradually become the mainstream method in related researches of learning analysis, which is helpful for in-depth exploration of the internal relations in learning data.

This study uses content analysis method to explore the application status of data mining technology in researches re-

lated to learning analysis, in order to provide reference for the future application of data mining methods in the field of learning analysis.

2. Study design

2.1. Data Source

In this study, the China National Knowledge Infrastructure (CNKI) database was selected as the source of literature research samples. First used the Boolean operation in advanced retrieval to search two keywords: "data mining" and "learning analysis", used the logical operator "AND" to connect the keywords, and checked the checkbox of education sector in "social scienceII album " in the catalog of literature classification, inall 339 related articles were retrieved. Through the screening of the retrieval results, non-applied research articles such as literature research, theoretical introduction and conference review were eliminated, and only literature related to the researches of learning analysis using data mining technology were selected as the research samples. Finally, 109 valid articles were obtained.

2.2. Research Methods and Tools

This study mainly adopts the content analysis method. Content analysis is a research method that objectively, systematically and quantitatively describes the research content (Feng, 2009). It usually carries out coding statistics and processing analysis on selected research units based on a certain coding framework, and effectively infers the communication content, so as to understand the essential facts and trends (Berelson, 1952). Therefore, this study uses content analysis method to explore the specific application of data mining technology in the field of learning analysis.

2.3. Research procedures

The content analysis framework in this study was developed by two researchers. The two researchers first designed the framework respectively based on the content analysis framework for related fields and analysis samples. Then made discussions on their own finished frameworks until a consistent framework was formed. The final content analysis framework is shown in Table 1.

After the content analysis framework was confirmed, two coders were selected to complete the coding work. First, two coders discussed the analytical framework and coding rules, and coded the literature samples independently on the basis of reaching agreement on the coding rules. After the coding was completed, the results of the two coders were compared, and the consistency coefficient of the coding results of the two coders was 0.894, which was relatively high. Then, the literature samples with inconsistent coding were discussed and recoded until the results reached a consensus.

3. Research Results

3.1. Research Topic

According to the statistics, among the 109 valid articles, 51 of them are theoretical exploration, 55 of them only carry out relevant model construction, and 3 of them both carry out theoretical research and complete relevant model construction. In theoretical exploration, some research discuss multiple topics at the same time, the amount of studies on each topic is shown in Figure 1. 34 articles (62.96%) focus on exploring the rules of teaching and learning, such as the exploration the impacts of learning patterns on learning effects (Wang, 2017), exploration on the relationship between the learning behavior and learning result (Chen & Song, 2019), analysis of the association between the students' feeling in class and learning result (Chen, 2018), analysis of the relationships between the students' characteristics, experiences, understanding and their satisfaction (Shu & Xu, 2014), etc. 14 articles (25.93%) have carried on the analysis of the learners, such as analysis of the online learning characteristics and learning situations of the learner groups (Chen, 2018), exploration of the differences of online learning behavior sequence among different types of learners (Jian, 2019), analysis of learners' learning status through mining relevant data (Lu & Yu, 2016), completion of the recognition of learners' emotional tendency (Liu, 2014), etc. At the same time also a small number of researchers carry on exploration on topics such as forecast (7.40%), self-adaptation and personalized learning (5.56%), evaluation (1.85%) and learning process monitoring (1.85%), for example, the realization of learners' achievement prediction based on the learners' behavior characteristics (He & Wu, 2016), and the completion of the active services' implementation of personalized education resources (Ding, 2016), etc.

Table 1. Content analysis framework

| The primary categories	The secondary categories	Level 3 categories
Research topics	The theoretical study	Exploring teaching and learning rules, self-adaptation and personalized learning, prediction, learner analysis
	Learner analysis model construction	Early warning, prediction, assessment, learner analysis, personalized education, others
Research objects	Preschool education	
	K12	
	Higher education	
	The remote education	
	Vocational education	
	Wisdom education	
	Others	
Data types	Data of Learners' characteristics	Gender, grade/age, student number, educational background, type, region, others
	Data on physiological and psychological responses during the learning process	Learning style, learning interest/preference, learning motivation, attitude and emotion, others
	Learning attention behavior	Length of study, frequency of study, quantity of study, login/attendance, choice of resources, others
	Autonomous behavior of learning	Browse, search, download, create resources, click, visit, thumb up, paths, others
	Learning interactive data	Comments and replies, posts, discussions, events, Q&A, others
	Learning outcome data	Test scores, final scores, questionnaire scores, notes, assignments, others
	The resource data	Resource information, resource characteristics
	Life data	Consumption, library borrowing, others
	Teacher data	
	Others	
Analysis methods	Statistical analysis	Descriptive statistics, regression analysis, correlation analysis, principal component analysis, others
	Visual analysis	Social network analysis, others
	Relationship mining	Association rules mining, sequence patterns mining, others
	Text mining	Content analysis, others
	Clustering and classification	Clustering, Bayes, Decision Trees, Neural Networks, others
	Recommendation algorithm	Collaborative filtering recommendation, others
	Others	
The data sources	Online data	Online learning platform, online learning management system, Web, others
	Offline data	
	Mixed form data	
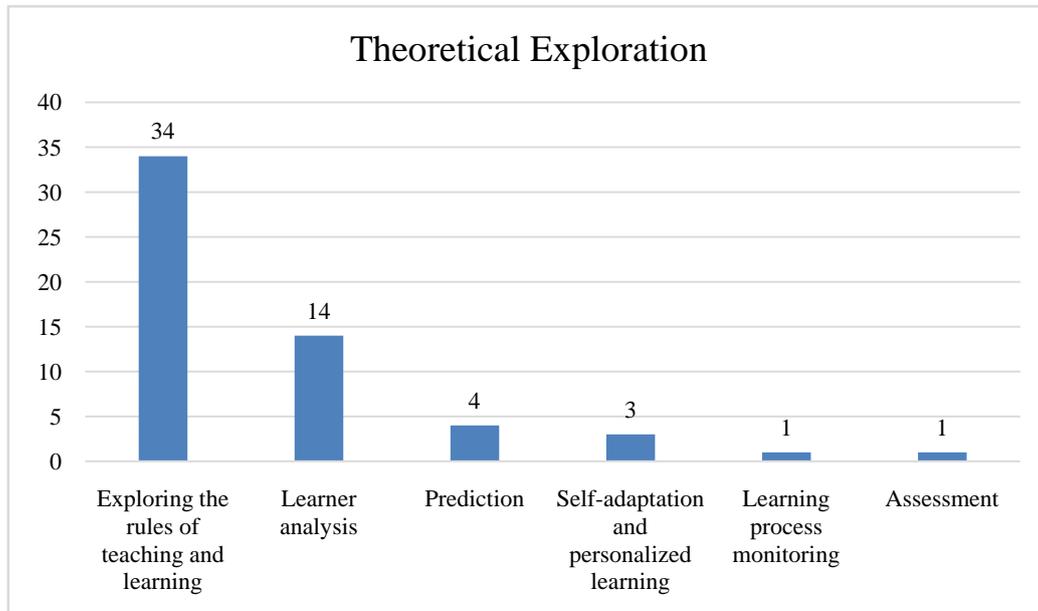
Figure 1. Quantity distribution of each topic of theoretical exploration.

In relevant research of model construction, the amount of each topic involved is shown in Figure 2. 32 articles (55.17%) built learner analysis models, for example, the construction of multiple concentric learning analysis model based on MOOCs (Hua & Zhang, 2014), the "6W" learning analysis framework proposed by integrating the six key elements of learning analysis (Gao, 2015), etc. In 11 articles (18.97%), evaluation models have been designed, such as building a multi-source and multi-dimensional model for comprehensive quality evaluation of students (Zhang & Qi, 2017). In addition, 10 articles have constructed prediction model and personalized education model, the percentage of both are 17.24%, for example, the construction of learning outcome prediction work model based on CIEO analysis thought (Mou, 2019), and the design of learning resources personalized recommendation model (Wang, 2018), etc. Besides that, 5 articles (8.62%) have built early warning model.
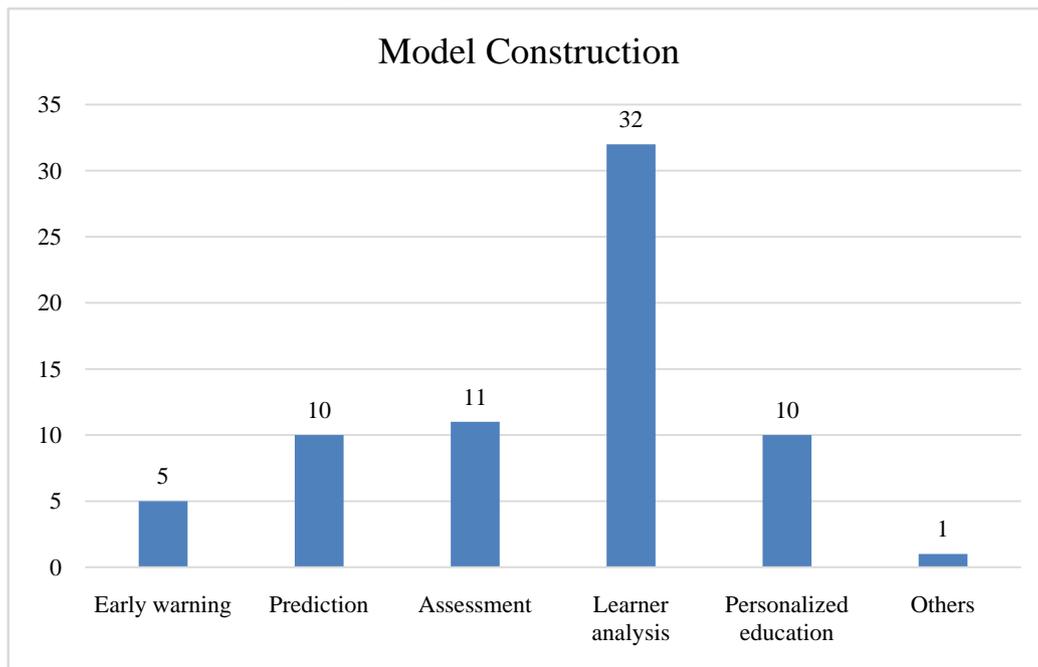

Figure 2. Quantity distribution of each topic of model construction.

3.2. Subjects

In 109 valid articles, only 67 described the specific research object, and the rest only pointed out that they were researching online learning. Among them, higher education received the most attention, accounting for 56.72%; 16.42% mainly focused on vocational education; in addition, some researchers also paid attention to K12 (10.45%), distance education (8.96%) and wisdom education (7.46%).

3.3. Data types

According to the statistics, each study involves at least three kinds of data, and the number of data types involved was shown in Figure 3. Among all valid articles, 83 (76.15%) collected learner's learning outcome data, that is, analyzed the relevant data generated by learners after the completion of learning, or used it as a learning result verification index, in which the final results, homework and test results were mainly concerned. 67 studies (61.47%) collected learners' interaction data, that is, data related to learners' communication and interaction in the learning process were collected, mainly focusing on learners' posting, reply and comment, discussion and so on. 63 studies (57.80%) analyzed the data of learners' attention behavior, that is, collected the relevant data of learners' attention to learning, which mainly focused on learning duration, learning frequency, login/attendance, etc. 62 papers (56.88%) collected the characteristic data of learners, that is, they focused on the characteristics and traits of learners themselves, mainly involving grade/age, gender and educational background, etc. 61 studies (55.96%) collected the data of autonomous behavior of learners, that is, focusing on the spontaneous learning behavior of learners in the learning process, mainly including browsing, visiting and downloading. 30 papers (27.52%) collected the physiological and psychological response data of learners in the learning process, that is, analyzed the relevant cognitive data of learners in the learning process, mainly involving learning style, learning interest/preference, attitude and emotion, etc. In addition, a small number of researchers collected data of teachers (5.50%), resource data (5.50%) and life data (4.59%) for analysis.
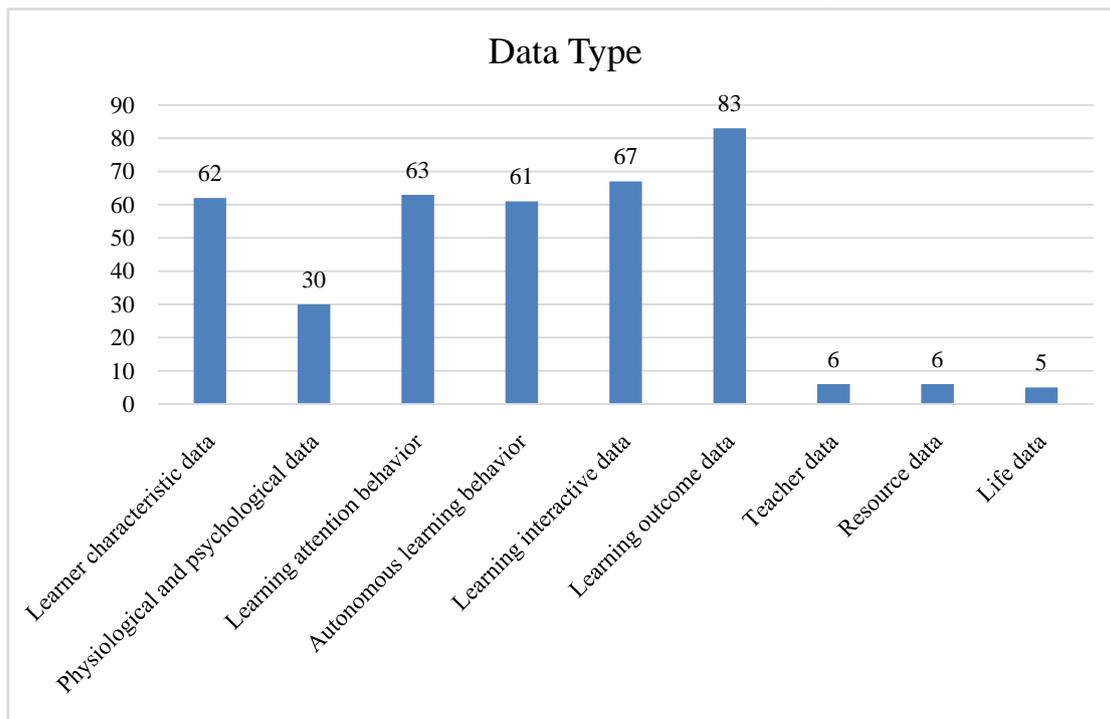

Figure 3. Quantity distribution of each data type.

3.4. Analytical methods

According to the statistics, a variety of methods are used comprehensively in most of the studies, and 2.80 methods are used on average in each paper. The number of analysis methods involved was shown in Figure 4. Among them, the most involved method is clustering and classification. The number of articles using this method is 69, accounting for 63.30%. It mainly includes clustering algorithm, decision tree and Bayes, etc., and is mostly used to classify according

to the characteristics of learners. 60studies (55.05%) used traditional statistical analysis methods, mainly involving correlation analysis, regression analysis and descriptive statistics. 46 papers (42.20%) involved the use of relational mining algorithms, mainly including association rule mining and sequence pattern mining to explore the relationship between different factors in the learning process. 24 papers (22.02%) used visual analysis methods, mainly including social network analysis and other types of visual presentation methods. 11 papers (10.09%) used relevant algorithms of text mining, such as text analysis, to analyze text data in the learning process. In addition, 9 articles (8.26%) used relevant recommendation algorithms, such as collaborative filtering recommendation, etc.
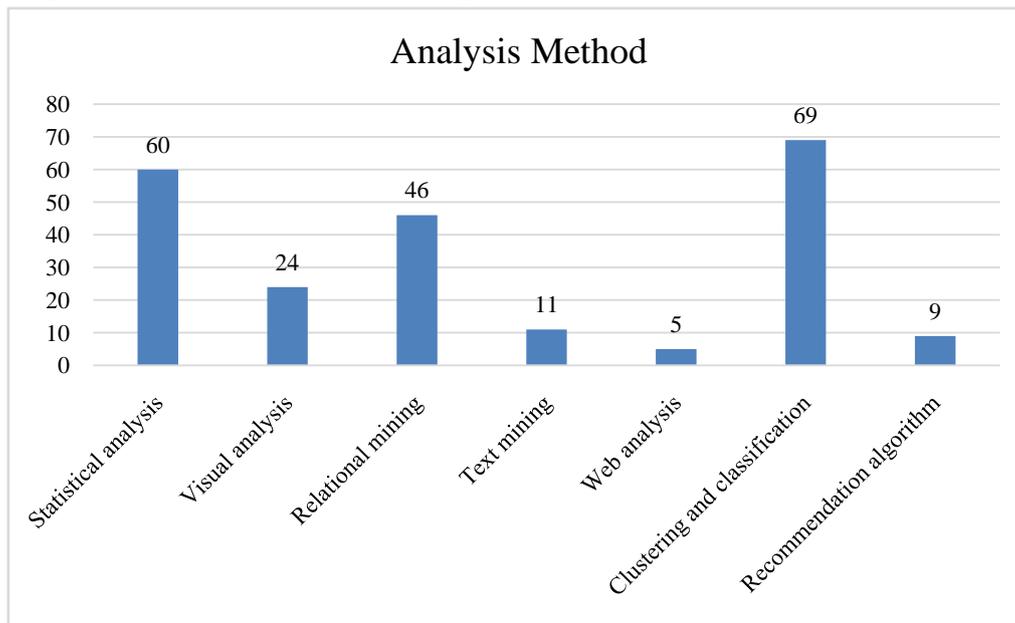

Figure 4. Quantity distribution of various analytical methods.

3.5.Data source

Among the 109 valid articles, 74.32% only selected online data, and the sources included online learning platform (72.84%), management system (22.22%), web log (3.71%) and website forum (1.23%). 11.00% only used offline data. 14.68% used mixed data.

4. Discussions

4.1. The research scenes and objects are relatively concentrated, and tend to be online and higher education

It is found that the research scenes of data mining technology in the field of learning analysis focus on online learning, which is closely related to the emergence of new learning methods such as MOOC, micro-class and mobile learning. Online platforms can record a large amount of data generated by learners during learning, which not only provides convenience for monitoring learners, but also promotes the realization of personalized education to a certain extent. Therefore, the study of online learning scenes has become the focus of learning analysis researchers.

At the same time, the analysis also shows that the target objects of relevant researches are mostly concentrated on higher education, and there are relatively few researches on K12 and preschool education. This may be because most college students have the basic information literacy of online learning, and have more opportunities to contact online learning, which is conducive to researchers to collect relevant data through online learning platform. The educational environment of higher education is also relatively open, where learners can generate a large amount of data in various aspects, and researchers need to use the corresponding data mining technology to deal with the huge data sources. However, in K12 and preschool education, learners are mostly in the traditional classroom learning environment, the data generated by learners are relatively limited, and the classroom behavior is difficult to record, which increases the difficulty for the promotion of relevant researches to a certain extent. Therefore, the application of data mining technology in K12 and preschool education learning analysis and its application mode still need to be further explored.

4.2. Research topics are gradually increasing, with equal emphasis on theory and practice

It is found that theoretical exploration and model construction are in a relatively balanced proportion in the relevant studies of learning analysis using data mining technology. Researchers not only pay attention to the exploration of relevant theoretical rules, but also pay attention to the construction of corresponding models to solve practical problems. In the study, analyzing learners and exploring the rules of teaching and learning are the hot spot of theoretical research and model construction. The support of data mining technology is helpful for researchers to analyze learners deeply from a wide range of data, to mine implicit features from explicit data, and to explore the objective laws in teaching and learning. Though prediction for the learners, the implementation of personalized education is not the current research hot spot, but with personalized learning continuous progress in the field of education, the realization of personalized education strategies, such as the prediction of learners' individual learning effectiveness and personalized resource recommendation, will become the focus of education researchers. In addition, some data mining techniques such as random forest and collaborative filtering recommendation also provide the possibility for the implementation of the corresponding strategies. Therefore, the prediction of learners and personalized education will become the research trend in the field of learning analysis.

4.3. Data types are increasingly wide and focus on the learning process

Study shows that the types of data collected are various. in addition to learning outcome data widely collected, behavioral data such as learners' interactive behaviors, attentiveness behaviors and autonomous behaviors in the learning process, as well as learners' own characteristic data are the hot spots for researchers to analyze. It can be seen that the support of data mining technology provides the possibility for the comprehensive analysis of diverse data in the learning process.

Compared with the above types of data, the collection of psychological and physiological data of learners is relatively small, and in most cases, the subjective scale was used for data collection. There is a lack of more objective psychological and physiological data collection, such as eye movement data. This may be because this type of data collection requires relatively expensive instruments, and it is difficult to collect a large number of students' data for data mining, so there are few relevant researches. How to balance the objectivity of data with the feasibility of research, and how to conduct data mining on psychological and physiological data to explore learners' emotional, cognitive and other psychological states still need further research.

In addition, the data of learning resources, teachers, and life data such as campus consumption and library borrowing in related management systems are also considered by some researchers. It can be seen that the introduction of data mining technology makes the types of data that can be analyzed increasingly extensive, which is conducive to the comprehensive analysis of the possible factors that affect learners' learning effectiveness.

4.4. Analytical methods are becoming more diverse, and data mining and learning analysis are increasingly closely linked

It is found through statistics that most researchers do not use a single analysis method in research, but a comprehensive use of multiple research methods. Although the traditional statistical analysis method is still widely used, but as a result of data mining methods provide convenience for large data processing and more deeply explore data internal rule, therefore it has become the mainstream of study analysis research method. The most widely used methods include clustering algorithm, association rule mining, decision tree, and so on. These methods can help us to cluster according to the characteristics of the learners (Hou, Li, Wang, &Li, 2011), explore the relationship between various influencing factors in learning and learning results (Chen, 2018), classification analysis of academic performance, etc. (Li, 2019).

In addition, other types of data mining technology in learning analysis researches are just beginning. For example, using text mining conducted semantic analysis of online discussion content (Lao, 2019), using web mining to analyze learners' web browsing history to explore their characteristics (Han & Lin, 2008). These methods enable automatic analysis of unstructured data, bring new research opportunities for researchers. For another example, social network analysis can visually present the demand relationship of learning resources in learning space (Yun & Liu, 2014), and collaborative filtering recommendation algorithm can be used to recommend personalized learning resources for students, etc. (Sun, 2018). It can be seen that the connection between data mining technology and learning analysis is becoming increasingly close. Data mining methods bring the possibilities for analyzing various data such as qualitative and quantitative data, and also promote the realization of multiple purposes such as exploring laws, forecasting and warning, and personalized education in learning analysis researches.

5. Conclusion

This paper analyzes the application status and trend of data mining technology in the field of learning analysis in China through content analysis method. It is found that the relevant research scenarios are mainly focused on online platforms and higher education. The main research topic is theoretical exploration and model construction of learner analysis and exploration of the rules of teaching and learning. The prediction of learners and personalized education are expected to become the trend. Data collection focuses on procedural and explicit data collection, and analysis of psychological and physiological data has received more and more attention. Analytical methods tend to be the comprehensive use of professional data mining methods.

The 9th International Conference on Learning, Analysis and Knowledge, held in the US in May 2019, highlighted the "inclusiveness" of analytics. The application of data mining technology in the field of learning analysis in China fully reflects this point. The diversity of data types and the integration of various analysis methods highlight the expansion of research "inclusiveness". I believe that future researchers will continue to consider enhancing research inclusiveness. For instance, expanding the scope of research objects and applying relevant data mining technologies to explore the learning rules in K12 and preschool education; Continuing to enrich data types, offline classroom behavioral data, unstructured data and physiological data can be more included in the research for a more comprehensive exploration. At the same time, the wide access to data also brings privacy leakage and other security risks. At present, relevant studies rarely mention the issues of "data privacy" and "data security". Therefore, researchers should pay more attention to this point and carry out researches on the premise of guaranteeing the legitimate rights and interests of the participants in the future.

References

Berelson, B. (1952). Democratic Theory and Public Opinion. Public Opinion Quarterly, 16(3), 313-330.

Chen, L. (2018). An Empirical Study on the Influence Mechanism of Teacher Trainees' Learning Behaviors in Online Education. Distance Education in China, 525(10), 37-45+81.

Chen, L. L. and Song, H. H. (2019). Analysis of Learning Behavior and Learning Effectiveness Based on MOOC Data Mining. Education Teaching Forum, 415(21), 59-60.

Chen, N. (2018). Students' Learning Data Mining Based on Double Teacher Cooperative Teaching Mode. MA.Sc. Thesis, Tianjin: Tianjin Normal University.

Chen, M. Y. (2018). Prediction Model Based on Association Rules and RBF Neural Network. MA.Sc. Thesis, Wuhan: Central China Normal University.

Ding, J. H. (2016). Personalized and Active Service Mode for Educational Resources. Ph.D. Thesis, Wuhan: Central China Normal University.

Feng, X. T. (2009). Sociological Research Methods. Beijing: China Renmin University Press.

Gao, J. (2015). Research on Construction and Application of Learning Analysis Framework. MA.Sc. Thesis, Changchun: Northeast Normal University.

He, C. K. and Wu, M. (2016). Analysis and Prediction of Learning Behavior of edX Platform Education Big Data. Distance Education in China, (06), 54-59.

Han, X. Q. and Lin, C. J. (2008). Analysis of Distance Learner Characteristics Based on Web Mining. Science & Technology Information, (33), 475-476.

Hua, Y. F. and Zhang, L. G. (2014). Research on Construction and Application of Multiplex and Concentric Learning Analytics Model Based on MOOCs. Journal of Distance Education, (05), 104-112.

Hou, Y. J., Li, Q., Wang, X. J., and Li, X. L. Effect Analysis of Distance Learners Action Based on K-means Clustering Algorithm. Journal of Beijing University of Posts and Telecommunications (Social Science Edition), (01), 108-113.

Jian, B. X. (2019). The Study on the Differences of Online Learning Behavior Sequences Based on Brain-based Learning Style. MA.Sc. Thesis, Changchun: Northeast Normal University.

Lao, C. Y. (2019). Implementation and Application of a Semantic Analysis Tool for Online Discussion in Network Learning Space. MA.Sc. Thesis, Wuhan: Central China Normal University.

Lu, L. S. and Yu Minghui. (2016). Learning State Analysis Method of Students Based on Outlier Detection. Computers and Modernization, (03), 35-40

Li, P. (2019). Management and Analysis of User Learning Data of MOOC Platform. Computer Education, (01), 145-149.

Lin, W. T. and Yang, N. (2017). Comparative Study on the Application of Learning Analysis Technology in Recent Five Years at

Home and Abroad. Journal of Fujian Radio & TV University, (05), 8-14.

Liu, Z. (2014). Research on Technologies of Emotion Recognition and Topic Mining for Course Comments. Ph.D. Thesis, Wuhan: Central China Normal University.

Mou, Z. J. (2019). Design and Empirical Research of Learning Outcome Prediction Based on CIEO Analysis from the Perspective of Learning Computing. e-Educational Research, 40(10), 68-75.

Sun, Y. H. (2018). Research on the Construction of Personalized Learning Environment Based on Big Data. MA.Sc. Thesis, Jinzhou: Bohai University.

Shu, Z. M. and Xu, X. D. (2014). Educational Data Mining and Analysis of College Students' Satisfaction from the Perspective of Learning Analysis. e-Education Research, (05), 41-46.

Wang, H. F. (2017). The Research on Students' Online Learning Behavior Based on Hadoop. MA.Sc. Thesis, Shanghai: East China Normal University.

Wang, H. N. (2018). Research on Learning Resources Personalized Service Strategy Based on Learning Analytics. MA.Sc. Thesis, Shanghai: East China Normal University.

Wang, M. X. (2013). Overview of data mining. Software Guide, 12(10), 135-137.

Yun, M. X. and Liu, H. (2014). A Regional Comprehensive Evaluation Model of Primary and Secondary School Academic Quality Based on Learning Analysis Technology. Journal of Shanghai Educational Research, (12), 24-26+45.

Zhong, X., Ma, S. P., Zhang, B., and Yu, R. Z. (2001). Overview of Data Mining. Pattern Recognition and Artificial Intelligence, 14(01), 48-55.

Zhang, Z. and Qi, Y. G. (2017). Construction of Multi-source and Multi-dimensional Comprehensive Quality Evaluation Model Based on Big Data. China Educational Technology, (09), 69-77+97.