Data Mining Based on CiteSpace Education Research Focus and Trend Analysis

Miaomiao Zeng*
School of Education, Zhaoqing University, Zhaoqing, China
*Corresponding author, e-mail: 2008020009@zqu.edu.cn

Abstract: *As the application value of educational big data becomes increasingly prominent, the development of educational data mining has been widely concerned. In this study, 1216 journal articles related to educational data mining were collected from CNKI database. By using information visualization software CiteSpace and using spatio-temporal knowledge graph and content knowledge graph analysis as the main research methods, this paper reveals and reflects the research hot spots and development trends of educational data mining in China, in order to provide reference for in-depth research, practical exploration and industrial promotion of educational data mining.*

Keywords: *Education data mining; Hot spots; Trend; CiteSpace*

Data mining (DM) is a process of discovering hidden patterns and knowledge from massive data through certain algorithms (Han & Kamber, 2001), which has been widely applied in banking, insurance, finance and other fields. With the development of educational informatization, the construction of smart campus and the exponential growth of big data in education, educational data mining (EDM) emerges at the historic moment, which aims to analyze the unique data generated in the educational environment to solve educational research problems (Baker & Yacef, 2009). Educational data mining bridges two disciplines: education and computational science, of which data mining and machine learning are sub-fields. EDM is defined by the education data mining community as: "Education data mining is an emerging discipline dedicated to developing new methods to explore unique and increasingly large amounts of data from educational environments and to use these methods to better understand students and their learning environments." In fact, EDM can also be understood as the application of DM in education big data, which is not only the embodiment of digital education research, but also the inevitable demand for the development of education informatization (Li & Fu, 2010). EDM, which is the integration of educational big data and data mining, has attracted more and more researchers' attention in recent years.

Research Program

Research Methods and Tools

Bibliometrics is a discipline that studies the distribution structure, quantitative relationship, change law and quantitative management of literature information by means of mathematics and statistics, and further discusses some structures, characteristics and laws of science and technology. CiteSpace is a widely used tool for bibliometric analysis. It is an information visualization tool developed by Professor Chen Chaomei of

Drexel University, which is specially used for academic literature analysis. It is suitable for multivariate, time-sharing and dynamic complex network analysis, which can detect hot topics and their evolution in a certain discipline or field. At present, it has been widely used to detect and analyze the changing trend of research frontier, the relationship between research frontier and knowledge base, and between different research frontier. CiteSpace's buttons mainly include Keyword, Cited Author, Cited Journal, Cited Reference, etc. Keywords are an important part and essence of academic papers, and their co-occurrence can keenly and directly reflect the hot spots and frontiers of research in a certain field (Yan, Zhu & Zeng, 2014).

The specific steps of CiteSpace analysis education big data research are as follows: Use CiteSpace's own data format conversion tool to convert CNKI literature exported to Refworks format into data format recognized by CiteSpace; Set the time span to 2002 to 2021, with an interval of one year. The thresholds (c, cc, ccv, where c is the citation frequency of references, cc is the co-citation frequency of references, and the co-citation coefficient of ccv references) were set as (2, 2, 20). Select Pathfinder's cut-link method to simplify Network structure and highlight important features, and use Cluster view-static and Show Merged Network visualization to present the final analysis map.

Research Objects

Select "Advanced Search" in CNKI, select "Topic" as "Education Data Mining", "Education & Data Mining", set the time from 2002 to 2021, and obtain a total of 1245 related literatures. After manual screening, reports, conference notices, documents, papers calling for contributions, preamble and so on were eliminated, a total of 1216 valid papers were obtained, including author, title, abstract, key words, author unit, reference and other fields.

Analysis of Research Results

Literature Time and Organization Distribution

The number of publications and citations can directly reflect the change of research popularity in a certain research field in a specific period of time, and is an important indicator to measure its development trend. To investigate the research results of educational data mining, this study makes statistics of literatures published from 2002 to 2021, as shown in Figure 1. As shown in Figure 1, from 2002 to 2010, the studies on education data mining increased year by year, but the overall studies were relatively few. Since 2014, the research literature on education data mining increased sharply, with rich research achievements and rapid development. This is attributed to the arrival of the era of big data, the application and development of educational data has been pushed into the "fast lane", especially the application of cloud computing, Internet of Things, mobile communication and other new information technology, which makes the collection of educational data more real-time, coherent and comprehensive. In addition, after 2014, as Maker education and STEAM education gradually entered primary and secondary schools, more and more intelligent products entered the education market, and data mining and learning analysis technologies were increasingly needed to solve problems in teaching.

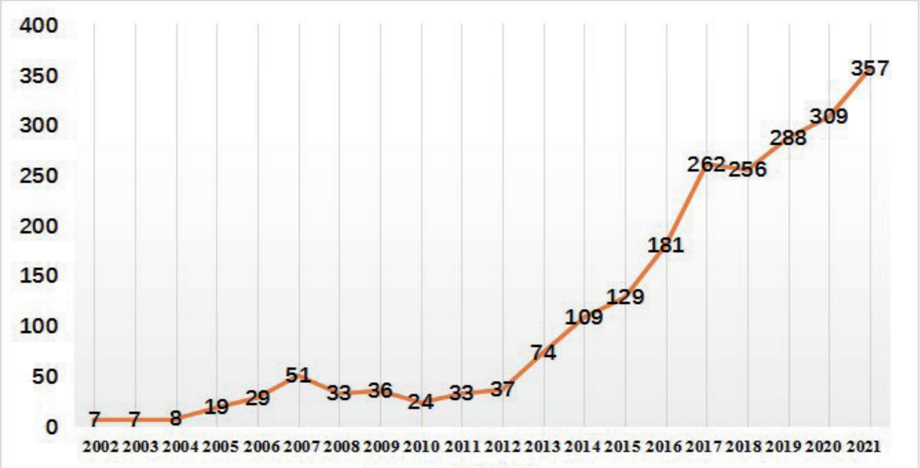


Figure 1. Literature chronology analysis of education data mining 2002-2021

Through the analysis of institutions and authors by CiteSpace, it was found that the co-occurrence network was very scattered, and the researchers appeared most frequently were Zhang Hai (11), Ding Guoyong (11), Zhu Zhiting (6), and Li Xin (4). The largest number of research institutions are East China Normal University and Northeast Normal University, but the research departments of these two schools are also very diversified, such as the Department of Educational Information Technology, School of Open Education, Department of Education, Shanghai Digital Education Equipment Engineering Center of East China Normal University. In Northeast Normal University, there are college of Media Science, College of Computer Science and Information Technology, Preparatory School for Studying in Japan, Office of Information Management and Planning, and College of Information Science and Technology. This also shows from another side that the research field of educational data mining is intersectional and dispersed, and the research results lack certain cooperation.

Word Frequency Analysis

From the perspective of knowledge theory, keywords with high centrality and frequency represent the common concerns of researchers in a period of time, that is, research hotspots. As a measure of the power of nodes, centrality reflects the importance of nodes in the network. The higher the co-occurrence frequency of keywords is, the higher the point centrality is, indicating that nodes are more important in this field. As shown in Table 1, keywords with the highest frequency in research literature include "data mining", "big data", "learning analysis", "educational data mining" and "association rules".

Table 1. High-frequency keywords

Serial number	**Keywords**	**Word frequency**	**Centrality**	**Serial number**	**Keywords**	**Word frequency**	**Centrality**
1	Data mining	359	1.21	12	Data analysis	14	0.02
2	Big data	110	0.19	13	MOOC	13	0.03
3	Study analysis	73	0.27	14	Personalized	13	0.03

4	Educational data mining	71	0.24	15	WEB data mining	12	0.02
5	Association rules	43	0.19	16	Online education	12	0.04
6	Distance education	42	0.27	17	Network education	12	0.03
7	Education big data	40	0.10	18	Teaching evaluation	11	0.01
8	Data mining technology	32	0.05	19	Apriori algorithm	10	0.03
9	Educational informatization	18	0.08	20	Learn analytical techniques	10	0.03
10	Data warehouse	17	0.11	21	Decision tree	9	0.06
11	Personalized learning	14	0.05	22	Machine learning	9	0.01

The keyword clustering function of CiteSpace can clarify the hot spots and development trends of a certain research field (Jiang, Zhao, Li & Wang, 2016). In the knowledge graph, circles represent keyword nodes, and the larger the circle is, the more frequently the topic appears. The color and thickness of node tree ring indicate the time period of appearance, that is, the thicker the color ring within the circle, the higher the frequency of appearance of the color in the corresponding year. The educational data mining literature data downloaded from CNKI database is processed, and the age of segmentation is 1 year. The source of clustering words is title, abstract, author information, key words, node type, etc., and the cutting line is set as the path detection algorithm, so as to obtain the educational big data literature clustering map (Feng, 2012). Analysis of keyword co-occurrence clustering results (see Figure 2) shows that there are 334 nodes and 677 links in the keyword co-occurrence network of educational data mining, and the overall network density is 0.0122. The centrality of "data mining" is the highest, which shows the importance and foundation of data mining theory and technology.

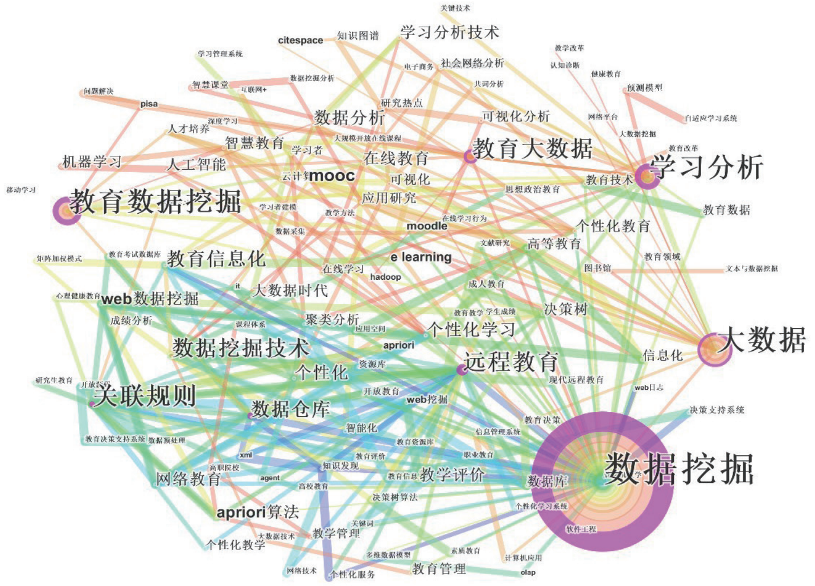


Figure 2. Keywords knowledge graph

Figure 2 shows that compared with institutional cooperation network, the structure and performance of keyword co-occurrence network have been greatly optimized and improved, but overall the structure of keyword co-occurrence network is still loose and the density is not high. In the future, relevant researchers should be required to do a good job in institutional research cooperation, and at the same time, they should maintain sufficient concentration on research topics and select appropriate topics for precise and in-depth research.

Analysis of Research Route

Based on the cluster graph, the time sequence graph of frontier keywords in education data mining is statistically analyzed by time segment in this study, which reflects the focus and change of fields of concern in the advancement and development of education big data, as shown in Figure 3. The development of educational data mining can be roughly divided into three stages: the first stage is the gestation period (2002-2006), which can be regarded as the technical development or ideological root of data mining, mainly involving the technical application of data mining and its proposal in education informatization, and later also involved in distance education. The second stage is the initial stage (2007-2013). As a special field, educational data mining was put forward, involving the application of distance education, education management, personalized education and learning, especially in the field of higher education and network education. The third stage is the development stage (2014-present), involving knowledge graph, machine learning, visual analysis, and the comprehensive emergence of various data mining technologies in performance analysis, talent cultivation, and education evaluation represented by Pisa large-scale evaluation.

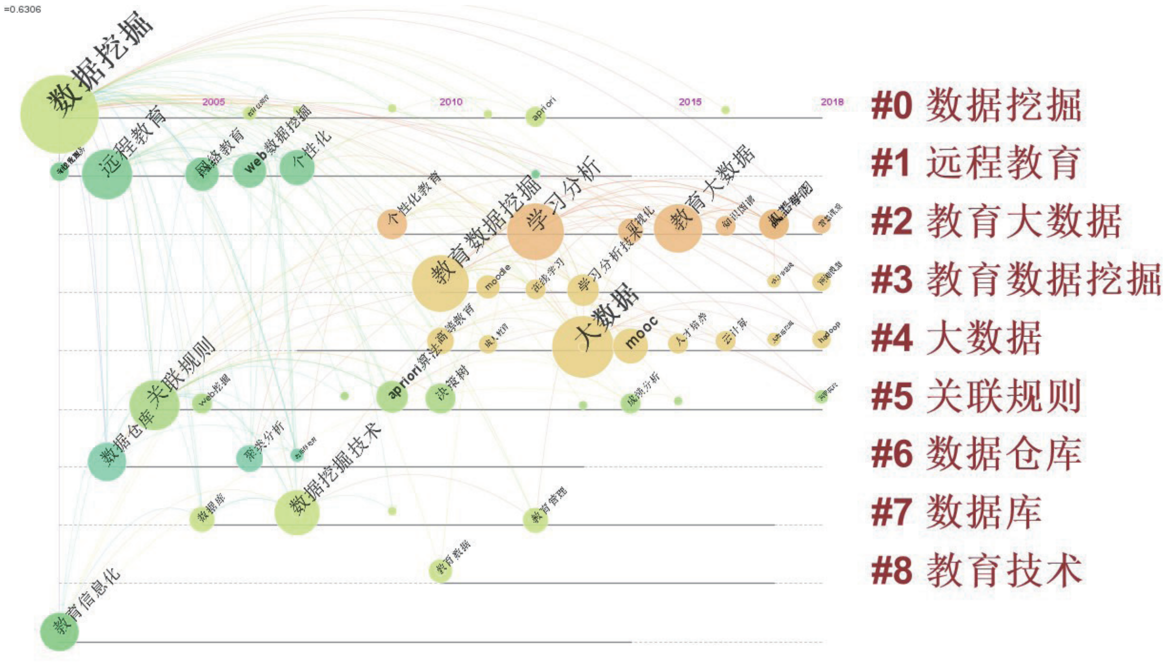
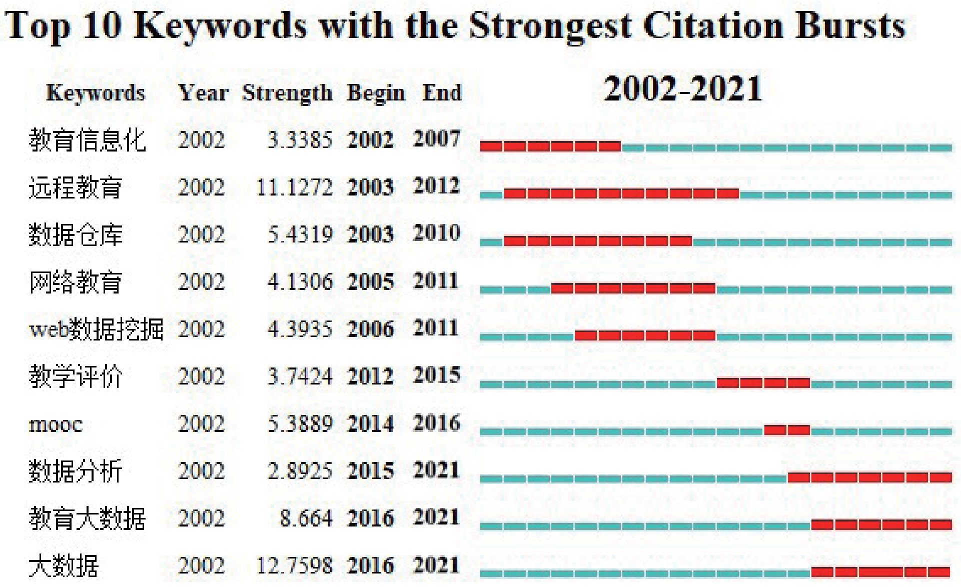


Figure 3. Time sequence atlas of frontier keywords

Research Trend Analysis

Mutants are words that appear more often or are used more frequently in a shorter period of time. According to the word frequency change of emergent words, we can judge the frontier and trend of the research field. According to CiteSpace related analysis, the emergent theme of educational data mining and its corresponding highlighting rate and cited history curve are obtained, as shown in Figure 4. "Education big data", "distance education" and "data analysis" are hot spots of education data mining research, among which "distance education" is mainly reflected in 2003-2013, "data analysis" in 2014-2018, and "education big data" in 2016-2021, and the research trend is increasing year by year. To a certain extent, the research frontiers of educational data mining in China are mainly embodied in educational big data, data analysis, distance and network education and other fields.



Top 10 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2002-2021
教育信息化	2002	3.3385	2002	2007	
远程教育	2002	11.1272	2003	2012	
数据仓库	2002	5.4319	2003	2010	
网络教育	2002	4.1306	2005	2011	
web数据挖掘	2002	4.3935	2006	2011	
教学评价	2002	3.7424	2012	2015	
mooc	2002	5.3889	2014	2016	
数据分析	2002	2.8925	2015	2021	
教育大数据	2002	8.664	2016	2021	
大数据	2002	12.7598	2016	2021	

Figure 4. Top 10 research hotspots

Analysis of Hot Spots in Educational Data Mining

Comparison and Consideration of Educational Data Mining and Learning Analysis

Educational data mining and learning analysis, which are closely related and overlapping, are popular fields in promoting teaching and learning. Although they have different origins and emphases, they share many common goals and concerns. Shu Zhongmei and Xu Xiaodong have made in-depth discussions on cross-field studies. For example, based on the perspective of educational data mining, factors affecting college students' satisfaction are obtained and meaning reconstruction is discussed from the perspective of learning analysis (Shu & Xu, 2014). The learning outcome evaluation model is constructed from both individual students and schools (Shu & Qu, 2014), and relevant factors based on student engagement model are identified by combining correlation analysis and data mining methods, and students' learning behaviors are classified (Shu, Xu & Qu, 2015). These discussions give full play to the advantages of modeling and discovery structure of educational data mining and promote the construction of meaning in learning analysis. At the same time, learning analysis is also used to collect learning traces to facilitate the discovery of evolutionary paths and calculation indicators of educational data mining algorithms. Liu Qingtang et al. (2017) defined the concepts and differences between learning analysis and educational data mining, and proposed that educational data mining could help solve the problems of insufficient data in learning analysis, and the application strategies of learning analysis could provide references for educational data mining.

Educational data mining and learning analysis, as the main application technologies of big data in education, should strengthen the exchange of their research results through their own journals and conferences in order to promote the development of educational practice and learning science. At the same time, competition and cooperation between the two communities can expand the number of researchers working in the field of big data in education, thus enhancing the application and impact of educational data mining and learning analysis through multidisciplinary collaboration in computer science, education, psychology, mathematics and other disciplines (Siemens & Baker, 2012).

Application Exploration of Distance Education Based on Education Data Mining

In educational data mining, extracting available information and constructing learner model are the boosters to solve key problems in distance education research. Among them, personalized learning support service and teaching interaction is one of the core contents of distance education, which is also the focus of researchers' attention. Jiang Qiang et al. (2016) emphasize the personalized adaptive learning become big data era digital learning, the necessity of the new normal, and refine the personalized the metacognitive and open learners adaptive learning model, autonomous learning mode and information visualization processing, to solve the problem of the network personalized learning, improve the effect of learners' learning, improve learning experience provides the reference. Feng Guier (2012) focuses on the role of data mining in distance education, uses association rules, cluster analysis and other methods to process online course data, respects individual differences and feeds back results to optimize teaching. Zhang Ting (2017), based on the development of modern distance education and personalized learning theory, uses data mining technology to mine learner information and designs a personalized learning system with learner model as the core, thus providing reference for subsequent intelligent application and dynamic update research. Therefore, educational data mining broadens the realization path of distance education development, and plays a great role in supporting personalized service, assisting learners to make choice of resources, learning diagnosis and feedback.

Research on Key Technologies of Educational Data Mining

In the context of "Internet + education", educational data mining technology is a means to gather and disseminate resources. It is also a kind of creativity, which is reflected in adapting to the changes of various educational scenes, dynamically mining hidden knowledge information, and forming new understanding and research. Educational data mining technology is actually a data mining technology used in the field of education. Feng Guier (2012) for a large amount of data online course learning, use all kinds of data mining technology in mining, such as visits from path analysis, clustering analysis of common characteristics, the association rules of statistical interest ratio, sequential patterns, predict learning behavior, make the online platform for the improved access and communication frequency increase, satisfaction and performance has been improved. Peng Ya et al. (2017) sorted out the distribution of educational data mining technologies and methods. The results showed that the commonly used technologies were prediction, relationship mining, clustering, statistical analysis and visualization, and collaborative filtering was less used. The focus of the research was that with the richer data and higher complexity of the technology, it was necessary to pay attention to the adaptability of mining task objectives and mining technology to avoid detours.

Researchers are more inclined to the application of data mining technology in network education, platform architecture, learning behavior analysis and other aspects, but the research on standardization and ease of use of technology is not mature, and the current problem of data "island" and technology "gap" still needs to be solved.

Study on Learning Effect and Academic Early Warning

Dynamic interaction in teaching and learning process, various types of data, such as learning engagement data, performance data, and so on, these can be used as education input, data mining is then prior rules automatically analyze the program output is obtained as a result, finally the results combined with education teaching analysis and application of the hot issues. As an important starting point for deepening the exploration of teaching practice, the study of academic performance prediction and academic warning is closely related to the learning activity process data in educational data mining. Chen Yijun et al. (2013) analyzed the performance characteristics of different network behavior groups based on the clustering algorithm in data mining, so as to discuss the influence model of students' performance and formulate effective strategies. Chen Zijian et al. (2017) adopted data mining and machine learning methods to jointly determine the factors influencing academic performance through correlation coefficient and information gain rate, build a classification prediction model and evaluate performance, and promote academic warning and learning prediction practice in online learning.

The general process of most research is: using data mining technology to extract and analyze various learning records and behavior information, build learner models, so as to determine the behavior attributes of learners, predict their learning performance, and provide directional content and targeted intervention for behaviors with warnings, so as to achieve accurate support for education and teaching, avoid risks and promote the all-round development of students. No matter in junior high school, senior high school or university, there is a considerable drop-out phenomenon. Studying the dropout factors of students can identify students at risk of dropping out in advance and carry out timely intervention to reduce the dropout rate. Different from the above studies on drop-out in traditional education, the high drop-out rate in online education has attracted the attention of many researchers.

Personalized Learning Services

Personalized learning services can provide students with the most appropriate learning resources, such as curriculum recommendation, personalized intervention, development of early warning system, etc. At present, there are mainly two kinds of researches on personalized learning service in the field of education data mining. One is personalized learning service based on recommendation system. Currently, personalized learning service based on recommendation system proposed by researchers mainly includes content-based recommendation algorithm, collaborative filtering and hybrid recommendation algorithm. For example, Zhu Tianyu et al. (2017) proposed a student-oriented collaborative filtering test question recommendation method, which can recommend appropriate test questions according to students' knowledge mastery. The other is personalized learning service based on data mining. Data mining methods for personalized learning services mainly include classification algorithm, clustering algorithm and association rules. In addition, some researchers are also using other technologies in personalized learning services.

Discussion on the Research Trend of Educational Data Mining

Guidance of Educational Data in the Learning Process

According to the analysis of mutant words, education big data and data analysis is one of the frontier and development trends. At present, distance education and network education are the most widely used fields of educational data mining. Achievement prediction and dropout research are mostly in higher education research. Educational data mining rarely involves the study of basic education and real classroom. This may be because the classroom teaching process in primary and secondary schools is more complicated and data acquisition is difficult. In May 2018, the Blue Book on The Development of Big Data in China's Basic Education (2016-2017) proposed six development trends and five challenges of big data in education, and put forward a series of development suggestions to the education industry, which is of great significance to promote the healthy development of China's education big data industry and education data mining. At the same time, with the construction of handheld devices and various smart classrooms, the teaching process will be recorded more completely and accurately, and the data analysis and mining of classroom teaching of basic education will also increase in the future.

Application of Deep Learning

At present, the research on educational data mining mostly uses several classical classification or regression algorithms to process data and find out the algorithm with the best performance. A few of them will adjust or improve the data mining algorithm according to the actual application scenarios. In fact, in addition to machine learning, deep learning has also achieved great success in image, speech recognition, natural language processing and other fields, but its application in the field of education is still not widespread. Deep learning technology with high prediction accuracy and no need to manually extract features will shine in educational data mining in the future.

Design and Application of Decision Support System and Adaptive Learning System

The algorithms, methods and technologies of educational data mining play an important role in the design and implementation of decision support system and adaptive learning system, and will promote the development and application of various service systems in the future. Decision support system consists of data warehouse, knowledge base, method library and man-machine interface. Technical improvement will be beneficial to play the synergistic effect of all parts, improve the system function, and provide strong support for school management and decision-making. Adaptive learning system is an effective way to support personalized learning and achieve differentiated teaching. It can transform and analyze multidimensional data and build models from social, emotional and metacognitive aspects, so that learners can master their own learning state and actively engage in deep learning. However, the current adaptive adjustment in recording, tracking, analysis, prediction, evaluation and other stages is not accurate, and further research is needed. In addition, the design and application of these systems need to apply more educational theories, aiming to provide guidance for content configuration, learning behavior evaluation and other aspects, jointly promote the realization of the significance of educational data mining, and avoid the embarrassing situation that practice is divorced from educational needs.

Application of Data Mining Technology in the Field of Teachers

At present, educational data mining rarely involves the study of teachers. This may be because it is difficult to acquire data about teacher development, which often requires questionnaires and other methods. Unlike student data, which is usually stored in educational administration management system, learning management system and other systems, it is convenient for direct application in education data mining research. As an important part of the field of education, it is expected that there will be more literature on teachers in the future.

Summary and Outlook

Existing studies on educational data mining have guided the development of data mining and educational practice activities to varying degrees, and promoted the field and height of the application of educational big data. At the same time, this study also found the deficiencies and difficulties in Chinese education data mining research by sorting out existing studies. Specifically, the following three aspects should be strengthened in the future research of Chinese educational data mining.

Enrichment and Improvement of Data Sets

At present, the research on education data mining in China is more focused on education field and theoretical level. First of all, there are few public data sets. (1) data sets usually contain the personal information of research objects, which should not be published based on academic ethics and norms. (2) data sets are usually valuable assets acquired by researchers at great cost of time and manpower. For the educational data mining community, the lack of high quality public data sets is one of the bottlenecks restricting its development. In addition, classifiers generated with only a small amount of data tend to have poor generalization ability, so data sets with larger sample size will be the development trend in the future.

Secondly, there is no unified data standard for data sets. The huge amount of information stored in the database, its expression form is complicated, which brings great trouble to the data preprocessing, so the standardization of data will be the focus of the future education data mining research.

Integration of Research Methods and Tools

At present, the research content of education data mining in China has attracted the attention of researchers from different disciplines such as computer science and education, but most EDM research results are published in educational journals rather than technical journals. Due to the lack of technical depth, educational data mining and learning analysis, as the main application technology of big data in education, have little cooperation. In the future, they should strengthen the exchange of research results through their own journals and conferences, so as to promote the development of educational practice and learning science. At the same time, researchers in the two fields are collaborating to enhance the application and impact of educational data mining through computer science, education, psychology, mathematics and other disciplines.

At the same time, the data mining tools used need certain domain knowledge. Data mining open source tools such as WEKA and Rapid Miner used in existing educational data mining literature all require certain knowledge in the field of mathematics or computer science, which are not very friendly to educators. Therefore, the development of open source tools suitable for the field of educational data mining is the engine to break down the development barrier of educational data mining.

Data Mining Based on China's Actual Education

At present, there are many evaluations and summaries of other countries' research achievements in China, but few researches on its own education status. Therefore, it is necessary to fully collect all kinds of educational data and use the technology of educational data mining to promote the "connection" and "intelligence" landing, and provide reference and basis for the teaching reform of schools at all levels and of all kinds, and bring new opportunities.

Funding

This paper is a phased achievement of the "13th Five-Year Plan" project of Education science of Guangdong Province "Practical Research on The Construction of" Artificial Intelligence + Professional Development of University Teachers "ecological Mechanism" [No. 2018GXJK206], and the "Recommendation research on Adaptive Path of Online Learning based on Knowledge Graph" of the higher education reform project of Guangdong Province in 2020（NO.612).

References

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techni-ques.* San Francisco: Morgan Kaufmann.

Baker, R. S., & Yacef, K. (2009). The state of educational datamining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17.

Li, T., & Fu, G. S. (2010). An analysis of the current situation and trend of Educational data mining research in China and abroad. *Modern educational technology,* 20(10), 21-25.

Yan, S. X., Zhu, N. B., & Zeng, Y. L. (2014). Key topics and evolution of Chinese curriculum research in the past 12 years: Visual analysis of knowledge map based on keywords co-occurrence in CSSCI Database from 2001 to 2012. *Global Education,* (3), 64-72.

Shu, Z. M., & Xu, X. D. (2014). The exploration and analysis of college students satisfaction from the perspective of learning analysis. *E-education Research,* 35(05), 39-44.

Shu, Z. M., & Qu, Q. F. (2014). An analysis of learning outcomes of college students based on educational data mining. *Journal of Northeastern University (Social Science),* 16(03), 309-314.

Shu, Z. M., Xu, X. D., & Qu, Q. F. (2015). Student engagement model and learning analysis based on data mining. *Journal of distance education*, 33(01), 39-47.

Liu, Q. T., Wang, Y., Lei, S. J., & Zhang, S. (2017). Student involvement model and learning analysis based on data mining. *Journal of distance education*, 35(03), 71-77.

Siemens, G., & Baker, R. (2012). Learning Analytics and Educational Data Mining: Towards communication and collaboration. Proceeding of *the 2nd International Conference on Learning Analytics and Knowledge*. New York, USA, 252-254.

Jiang, Q., Zhao, W., Li, S., & Wang, P. J. (2016). Research on personalized adaptive learning—The new normal of digital learning in the era of big data. *China Educational Technolog*, (02), 25-32.

Feng, G. E. (2012). Application of data mining technology in Distance Education. *Modern educational technology,* 22(12), 96-98.

Zhang, T. (2017). *Research on personalized learner model in Modern Distance Education* [Doctoral Dissertation, Jiangnan University].

Peng, Y., Yu, C. B., & Zhang, X. (2017). Research on the application of educational data mining technology. *China Educational Technology & Equipment,* (18), 1-5, 13.

Chen, Y. J., & Yin, L. (2013). Research on student achievement influence model based on data mining. *Modern educational technology,* 23(01), 94-96, 93.

Chen, Z. J., & Zhu, X. L. (2017). Research on online learning achievement prediction based on educational data mining. *China Educational Technology,* (12), 75-81, 89.

Zhu, T. Y., Huang, Z. Y., & Chen, E. H. (2017). Personalized test question recommendation method based on cognitive diagnosis. *Chinese Journal of Computers*, (01), 178-1.