A Data Mining Based Method for Identifying the Mismatch of Educational Resources in Regional Colleges and Universities

Yan Liu(✉), Sun Hui, and Sen-wei Wang

Dalian University of Science and Technology, Dalian 116052, China

Abstract. When identifying the mismatch of educational resources in regional colleges and universities, there is a certain mismatch between the elements of resource mismatch and the measurement, which leads to the problem that the actual recognition speed is too low. This paper constructs a data mining based method to identify the mismatch of educational resources in regional colleges and Universities. After determining the subject of collection, mining the data of educational resources in Colleges and universities, using all elements of resources to proxy variables, controlling the matching parameters between elements and measures, measuring the mismatch of educational resources, integrating similar teaching resources, establishing mismatch recognition rules, and finally completing the identification of educational resources mismatch. This paper simulates the regional university resource environment, selects the information science, information work, discipline field and related literature of CNKI as the mismatched educational resources. Experiments are used to verify the effectiveness of this method. The results show that the mismatch identification method designed in this paper can identify the mismatched resources quickly maximum.

Keywords: Data mining · Regional universities · Educational resources · Mismatch

1 Introduction

For a long time, University Library, as an academic, scientific research and service institution, has played an important role and accumulated a lot of information. With the development of science and technology, University Library urgently needs to use advanced technology and means to process all kinds of data information in order to improve service quality and resource allocation efficiency, and to promote university library to continuously improve its own business work and management level. Although the application of data mining technology in domestic library is still in the initial stage, with the development of information technology and the change of user demand, data mining technology will combine with the work of Library and play a great guiding role. The service object of university library is relatively fixed, mainly for the different needs of university teachers and students in learning, scientific research and work. From

W. Fu et al. (Eds.): eLEOT 2021, LNICST 390, pp. 291–302, 2021.
https://doi.org/10.1007/978-3-030-84386-1_24

the form of service, passive service is more than active service [1]. From the level of service, although the scope is relatively fixed, there are many levels, such as the level of undergraduate and graduate students, as well as the scope of research, the scope of interest and other aspects [2]. Therefore, the service of university library should be targeted and professional.

With the development of data mining technology, foreign countries began to study its application in the field of library, and applied the actual results to practice, and its application mode has become increasingly mature. By the end of 1990s, with the rapid development of the Internet, the construction of digital library, the construction of knowledge and information resources under the network environment, intelligent retrieval, interlibrary loan and other new topics have been emerging, and the development and research based on the application technology of digital library has become more detailed [3]. Foreign scholars have made some progress in the application of data mining technology. With the rapid application of the Internet, library digitization, knowledge and information resources construction, intelligent retrieval, interlibrary loan and so on have become research hotspots and made great achievements [4]. Digital library application technology has also made rapid development and wide application.

The innovation of this article is based on data mining technology. The elements in the resource are used to represent variables. Data mining technology is used to collect educational resources in colleges and universities. Resources are used for proxy variables. The adaptation parameters between the element and the measure are controlled. Similar teaching resources are integrated. The identification rules are established.

2 Identification Method of Regional University Education Resources Mismatch Based on Data Mining

2.1 Mining University Education Resource Data

Before collecting the data of university education resources, first determine the subject of collection, select the university resources of the same specialty as the collection object of behavior data, and collect the educational resources of colleges and universities according to the following processing process. The collection process is shown in Fig. 1 below:

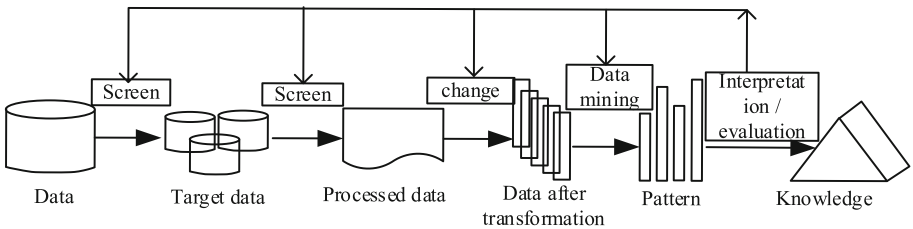

Fig. 1. Acquisition process

As shown in Fig. 1 above, firstly, according to the teaching management system of colleges and universities, collect and record the field of educational resources in Colleges and universities, formulate a data screening standard, sort out the data sets to be analyzed, and then select the data sets related to data mining to narrow the collected data range and enhance the mining quality of behavioral data [5]. Then this part of the data is preprocessed to remove the noise in the original collected data, filter the redundant and invalid data, and use the criterion function of K-average algorithm to fill in the missing data

$$E = \sum\nolimits_{i=1}^{k} \sum\nolimits_{q \in D_i} |q - n_i|^2 \tag{1}$$

In the above formula, E is the sum of the mean square deviation of all behavior data in the data set and the corresponding cluster center, q is the given data object, n_i is the mean value of clustering, q, n is multidimensional. The data filled with formula (1) is combined with the previous semi-structured data to be processed into structured data [6], and the naive Bayes classification method is used for classification. The classification process is as follows:

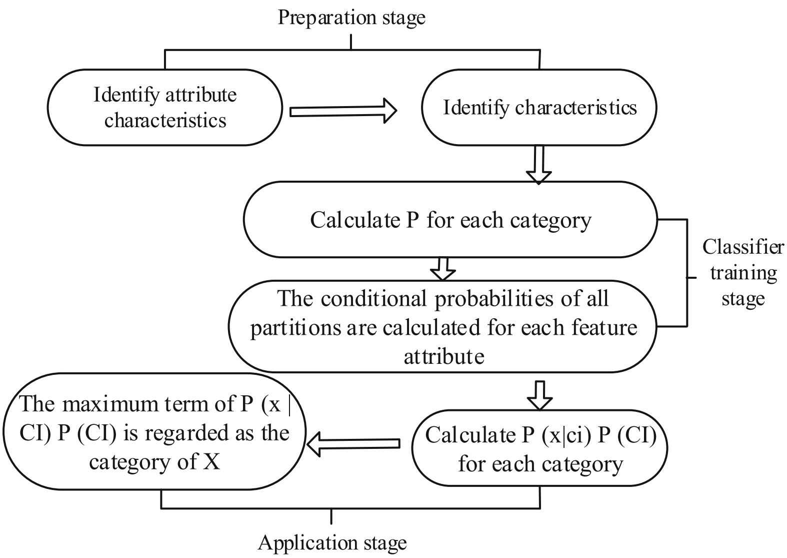

Fig. 2. Classification process

As shown in the above Fig. 2, first assume that each feature Y of the sample is uncorrelated and in an independent state C. Set it as the set of behavior data training tuples and corresponding class labels. Each tuple is represented by a dimensional vector $B_1, B_2, \ldots B_n$, in which the attributes of teaching resources measure the tuple C_i. Under one condition, the predicted tuple belongs to a class, so that the C_i has the highest posterior probability [7], and there is the following relationship:

$$P(C_i|Y) > P(C_j|Y) \quad 1 \le j \le m, j \ne i \tag{2}$$

From the above formula, $P(C_i|Y)$, the results can be obtained:

$$P(C_i|Y) = \frac{P(C_i)P(Y|C_i)}{P(Y)} \tag{3}$$

In the above formula, $P(Y)$ is the fixed value in the known data, the above collected and classified data are retained, and then the feature extraction processing of the collected and classified college education resources [8] is used to measure the mismatch of educational resources.

2.2 Measuring the Mismatch of Educational Resources

To measure the mismatch of educational resources, the total elements of resources [9] are used as surrogate variables to calculate the stock of educational resources in regional colleges and universities.

$$K_{it} = \frac{K_{it-1} + I_{it}}{D_{it}} \tag{4}$$

Among them, K_{it} represents the resource stock of regional colleges and universities in the region, I_{it} represents the introduction of educational resources by colleges and universities in the region, and D_{it} represents the configuration parameters. In order to eliminate the inconsistency of resource estimation among different university individuals, a resource update function is constructed

$$Y_{it} = \frac{A_{it}K_{it}^{\alpha}}{L_{it}^{\beta}} \tag{5}$$

Among them, Y_{it} refers to the number of resources replaced in Colleges and universities, A_{it} refers to the updating parameters of educational resources, L_{it}^{β} represents the number of types of educational resources, β and α respects represents the renewal elasticity of educational resources. In order to further estimate the elements of educational resources, the above calculation formulas (4) and (5) are combined to calculate the logarithm, which can be expressed as follows:

$$\ln Y_{it} = \frac{a_{it} + \alpha \ln K_{it}}{\beta \ln L_{it}} \tag{6}$$

Among them, the logarithm a_{it} of represents, and the meaning of other parameters A_{it} remains unchanged. The regional distribution of colleges and universities and the distribution of their educational resources have the characteristics of non degradation, and the relationship between the two in different regions or countries is not completely the same. A significant positive relationship between the mismatch rate and the size of regional colleges and universities was found, which shows that the universities in large regions tend to have higher mismatch rate of resources. OP covariance [10] can be a good measure of resource mismatch. Therefore, the size of the university area is used as a surrogate variable for the size, and use OP decomposition method to calculate the covariance difference between university education resources and mismatch. The op covariance is used as a measure of resource mismatch. To a large extent, it makes up for the possible defects in the mismatch method in the literature. The static OP decomposition method is adopted, which does not consider the dynamic changes of university education resources, that is, it does not take the new form of educational resources into consideration. This paper mainly decomposes the mismatch rate of each university in the same period

$$\Phi_t = \overline{\varphi_t} + \sum_i (S_{it} - \overline{S}) \tag{7}$$

Among them, $\overline{\varphi_t}$ represents the renewal rate of educational resources, S_{it} represents the share of educational resources in Colleges and universities, and $\overline{S}$ represents the average value of educational resources in Colleges and universities. Taking time as an independent variable, the change of university education resources mismatch is calculated, as shown in Fig. 3.

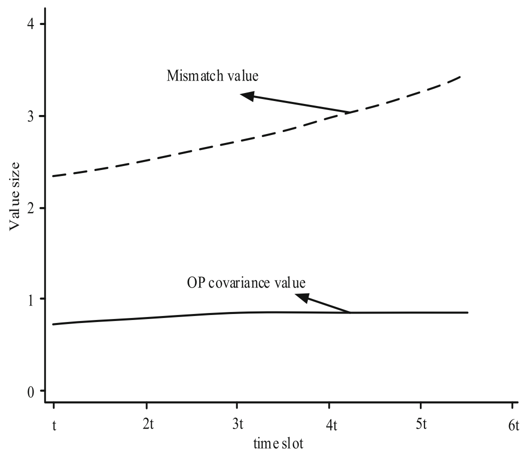

Fig. 3. The mismatch of educational resources in Colleges and Universities

As shown in Fig. 3, with the increase of OP covariance difference [11–13], the mismatch value of educational resources is also increasing, and the mismatch value almost reaches the value of 4 at the end of the statistical cycle. According to the numerical changes caused by the above mismatching process, the complexity of university education resources is defined.

2.3 Establish Mismatch Recognition Rules

Before establishing the recognition rules, the similarity teaching resources are integrated [14], and the similarity matrix is constructed by similarity evaluation method

$$S_D(d_i,d_j) = \begin{cases} v * SS_D(d_i, d_j), if d_i d_j \in DT \\ GS_D(d_i, d_j), otherwise \end{cases} \tag{8}$$

Among them, DT represents the set of mismatched information data and v is the weight parameter. Suppose the mismatched resource is g_i. The weights g_j between the two mismatched resources are calculated

$$LLS_N(g_i, g_j) = \frac{LLS(g_i, g_j) - LLS_{\min}}{LLS_{\max} - LLS_{\min}} \tag{9}$$

Where, $LLS(g_i, g_j)$ denotes the weight between mismatched resources, and represents the minimum $LLS_{\min}$ and $LLS_{\max}$ maximum log likelihood scores of mismatched resources, respectively. Then the functional similarity [15] between the two genes can be expressed as follows:

$$FS_G(g_i, g_j) = \begin{cases} 1, & g_i = g_j \\ 0, & e(g_i, g_j) \notin HumanNet \\ LLS_N(g_i, g_j) \in HumanNet \end{cases} \tag{10}$$

Among them, $e(g_i, g_j)$ represents the boundary value between the regional university education resources g_j and. using the similarity calculation results, the final mismatch identification results are for g_i mulated, and the module information recognition rules are finally calculated, as shown in Table 1.

Table 1. Identification rules established

Rule name	Resource name	Number
Recognition rule 1	Learning video	127
Recognition rule 2	Open class	48
Recognition rule 3	TV open class	145
Recognition rule 4	Video open class	40
Recognition rule 5	Excellent course	32
Recognition rule 6	Video conference system	11
Recognition rule 7	Electronic library resources	127
Recognition rule 8	Online teaching course	62
Recognition rule 9	Documentary of teaching resources	139
Recognition rule 10	Teaching materials	4

After establishing the recognition rules in the Table 1, the rules in Table 1 are used to realize the recognition of the mismatch of educational resources of regional universities [16, 17], and finally complete the research on the mismatch recognition.

3 Simulation Experiment

3.1 Experimental Preparation

The spider web system is prepared as the experimental environment of mismatch identification method. In the actual deployment, the system is refined into front-end processing, distributed storage and recognition display. In order to deal with the massive network traffic in large-scale network environment, the front-end processing part is further divided into: front-end packet capture machine and front-end processor. The front-end packet capturing machine is mainly responsible for capturing data packets in the network, and sending multiple packets with the same quintuple to the same front-end processor; The front-end processor is responsible for the processing of data packets, including flow table management, IP fragmentation reorganization, TCP reorganization, etc. at the same time, according to the traffic characteristics of complex applications, the statistical characteristics are extracted based on influenza knowledge model, and the complex application traffic is identified by saimmf method, and the logs are sent to the distributed storage subsystem.

Distributed storage uses multiple servers to build a distributed storage cluster based on Hadoop framework to ensure the storage reliability of data; at the same time, spark and shark are used to achieve efficient distributed computing. Analysis and display is divided into data analysis machine and data display machine. Among them, the data analysis machine is responsible for mining massive log data, modeling from multiple dimensions using network flow field method, and constructing private network. Data display machine is responsible for displaying the results after mining, and its display content mainly includes statistical information and real-time animation. The deployed verification experiment environment is shown in Fig. 4.

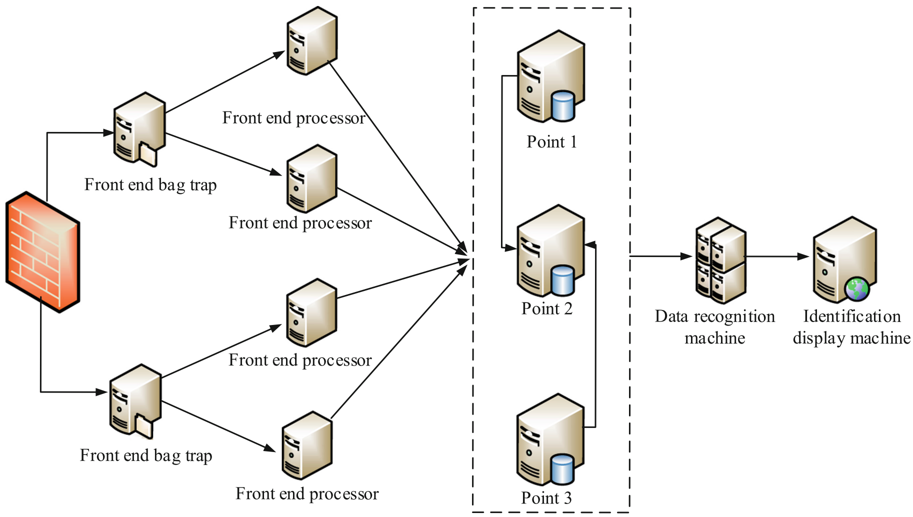

Fig. 4. Test environment of experimental deployment

In the test environment as shown in Fig. 4, the degree of association between knowledge nodes is the representation of resource aggregation, and the aggregation group is the collection of nodes with strong connection. The higher the unit requirements of the aggregation group, the higher the aggregation depth, and the more likely the knowledge nodes within the aggregation group will converge. The research sample of this paper is from CNKI database. The subject field of "information science and information work" subordinate to "library, information and digital library" in the database is selected, and the publication time of the sample is 2017–2019 (Due to the time lag of articles included in CNKI, the data was obtained in 2017. To ensure the accuracy of the data, the 2017 data was not selected as the research sample.). Based on the above conditions, a total of 12815 papers were retrieved, and 11288 papers were obtained after the critical reports and index articles were screened out. After the above resources are used as the educational resources of regional colleges and universities, the mismatch recognition methods in literature [4], the mismatch identification methods in literature [6] and the mismatch recognition methods designed in the paper are used to compare the performance of the three methods.

3.2 Results and Analysis

Based on the above experimental preparation, the size of educational resource packet is set to 64 bytes. Since it is impossible to manually label the complex application network traffic captured in the real network, it is difficult to accurately evaluate the identification performance of spider web system for specific complex applications. In order to solve this problem, manual access is used to generate network traffic for specific applications in the laboratory environment, and tcpdump tool is used to capture the corresponding application data packets. After that, the network traffic is injected into the spider web system by playback to identify and record the identification results. In the above experimental environment, control the wire speed of 100Mbps, 200Mbps, 500mbps and 800Mbps to send university education resources to the mismatch module. Finally, the number of data packets identified by the three mismatch identification methods is calculated and the recognition rate is calculated. The results are shown in Fig. 5.

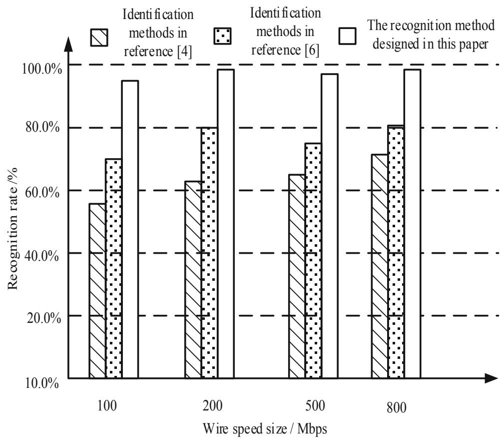

Fig. 5. Recognition rate results of three recognition methods

It can be seen from the experimental results shown in the above figure that under the same experimental environment, different recognition methods are controlled to identify educational resources with different linear speeds. According to the experimental results shown in the above figure, the average recognition rate of the recognition method in reference [4] is 60% on the left and right under four thread conditions, and the final average recognition rate of the identification method in reference [6] is about 70%, while the recognition rate designed in this paper is The average recognition rate of each method is about 97%. Compared with the mismatch recognition methods in the two literatures, the final recognition rate of the proposed method is the highest.

Keep the above experimental environment unchanged, set the data set of educational resources mismatch, take the number of TCP flows and UDP flows in the resources as the setting objects, and set the resource test data set as shown in Table 2.

Table 2. Set of test data sets

Data set	Number of TCP streams	Number of UDP streams
Resource dataset 1	176381	675531
Resource dataset 2	3194	547444
Resource dataset 3	83474	264107
Resource dataset 4	219474	1405242
Resource dataset 5	14425	1876533
Resource dataset 6	172291	451302

The test data set shown in the table above is taken as the processing object. Defining three mismatch methods to meet the storage requirements of resource log is a precise identification process. The number of accurate identification resources is obtained by counting the three identification methods, and the final accuracy rate of the three mismatch identification methods is calculated. The accuracy results are shown in Fig. 6.

From the accuracy results shown in the figure above, the three identification methods show different accuracy results after changing the resource allocation process with different line speeds. According to the precise value shown in the figure above, the average accuracy value obtained by the identification method in reference [4] is about 75%, and the accuracy value is the smallest. The average accuracy rate of the identification method in reference [6] is about 85%. The average accuracy of the proposed method is about 97%. Compared with the two methods in literature, the recognition method designed in this paper has the highest recognition accuracy, which meets the accuracy requirements of Library mismatch recognition.

Keep the above experimental environment unchanged, take the number of resources identified by the three mismatch identification methods in unit time as the comparison index, and the data in the test data set prepared in the above table as known. Count the final processing time of the three mismatch identification methods, and calculate the number of resources identified by different recognition methods in unit time. The experimental results are shown in Table 3.

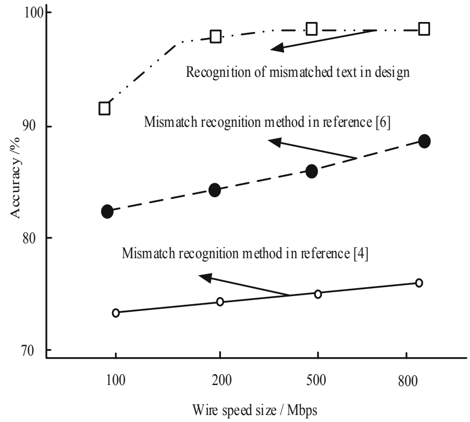

Fig. 6. Accuracy results of three identification methods

Table 3. Identification speed results of three resource mismatch methods

Data set Resource dataset 1	Recognition speed/(bar/sec)		
	Identification methods in reference [4]	Identification methods in reference [5]	Identification methods in reference [6]
Resource dataset 2	47.4	87.4	125.2
Resource dataset 3	49.5	86.5	158.7
Resource dataset 4	51.4	89.2	125.9
Resource dataset 5	51.1	93.9	134.6
Resource dataset 6	42.8	93.7	158.1
Data set	49.8	99.6	158.6

According to the results of recognition speed shown in Table 3, with the same resource test data set as the experimental object, under the control of the three identification methods, the recognition speed of the identification method in reference [4] is about 48 bars/second, and the speed of resource identification is relatively small. The recognition speed of the identification method in reference [6] is about 91 pieces/second, and the actual recognition speed is larger. The recognition speed is about 143. Compared with the two methods in literature, the recognition method designed in this paper has the fastest recognition speed. According to the above experimental results, the recognition method designed in this paper has the maximum recognition rate, the maximum accuracy rate of the recognition process, and the actual recognition speed, which is suitable for the application in the process of university education resources mismatch recognition.

4 Conclusion

With the advent of the era of big data, the amount of information is increasing rapidly. For the resource construction of digital library, resource aggregation has become an urgent problem. As an effective means of content aggregation, aggregation technology plays an important role in organizing massive resources and effectively establishing the connection between resources. With the support of data mining technology, the construction of a regional university education resources mismatch identification method can improve the lack of slow identification of resources in the literature, and provide certain theoretical basis for future research work.

References

1. Ding, G., Cheng, J.: Modeling of college students' academic performance based on educational' data mining. Heilongjiang Res. High. Educ. **38**(02), 76–81 (2020)
2. Wang, R., Wang, C., Li, J.: An intelligent divisional green tide detection of adaptive threshold for GF-1 image based on data mining. Acta Oceanol. Sin. **41**(04), 131–144 (2019)
3. Yin, Z.: Modeling and analysis of book borrowing flow in university library based on data mining. Microelectron. Comput. **35**(11), 95–99 (2018)
4. Zhong, R., Wang, H.: Data mining based specific data query technology for cloud computing management system in university. Mod. Electron. Tech. **41**(02), 130–132 (2018)
5. Liao, C., Bai, X., Xu, M.: Correction of space atmospheric model based on data mining method. J. Beijing Univ. Aeronaut. Astronaut. **44**(12), 2628–2636 (2018)
6. Liu, B.-P., Fan, T.-C., Yang, H.: Research on application of early warning of students'achievement based on data mining. J. Sichuan Univ. (Nat. Sci. Ed.) **56**(02), 267–272 (2019)
7. Chai, Y., Lei, C.: Survey of online learning behavior research applying data mining technology. Appl. Res. Comput. **35**(05), 1287–1293 (2018)
8. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
9. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
10. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
11. Lu, L., Zhou, J.: Research on mining of applied mathematics educational resources based on edge computing and data stream classification. Mob. Inf. Syst. **2021**(7), 1–8 (2021)
12. Bhat, M., Zaman, M., Butt, M.: An intelligent prediction system for educational data mining based on ensemble and filtering approaches. Procedia Comput. Sci. **2**(167), 1471–1483 (2020)
13. Prada, M.A., Dominguez, M., Vicario, J.L., et al.: Educational data mining for tutoring support in higher education: a web-based tool case study in engineering degrees. IEEE Access **8**, 212818–212836 (2020)
14. Ang, L.M., Ge, F.L., Seng, K.P.: Big educational data & analytics: survey, architecture and challenges. IEEE Access **PP**(99), 1 (2020)
15. Jung, I., Lee, J.: A cross-cultural approach to the adoption of open educational resources in higher education. Br. J. Educ. Technol. **51**(1), 263–280 (2020). https://doi.org/10.1111/bjet.12820

16. Nguyen, A., Gardner, L., Sheridan, D.P.: Data analytics in higher education: an integrated view. J. Inf. Syst. Educ. **31**(1), 61–71 (2020)
17. Karthikeyan, V.G., Thangaraj, P., Karthik, S.: Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. Soft Comput. **24**(5) (2020)