Big Data-Driven Automated Engagement Quantification Model for Language Learning

*1st Lihong Wang
School of Basic Education
Beijing Polytechnic University
Beijing, China
*wanglihong@bpi.edu.cn

2nd Runfen Yang
School of Foreign Languages and Cultures
Beijing Wuzi University
Beijing, China
yangrunfen@bwu.edu.cn

3rd Weijie Gou
School of Automotive Engineering
Beijing Polytechnic University
Beijing, China
gvjie@126.com

2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) | 979-8-3315-2228-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/AINIT65432.2025.11035764

***Abstract*—To address the limitations in accuracy associated with conventional assessment methods for English learning engagement, this study proposes a novel big data-driven automatic computation model. The framework incorporates a systematic data preprocessing phase utilizing advanced big data analytics to ensure dataset integrity. A multidimensional online learning behavior analysis index system was established through entropy weight method and analytic hierarchy process (AHP) for precise indicator weight allocation. Particularly, we introduced a hierarchical clustering algorithm combined with Pearson correlation matrix analysis to categorize behavioral correlations into distinct feature groups. These grouped characteristics were subsequently integrated into an ensemble learning architecture to construct the final engagement quantification model. Experimental validation demonstrates that the proposed model achieves a 40% improvement in measurement accuracy compared with traditional approaches (p<0.01), while reducing computational latency by an average of 12 minutes per assessment cycle. The model effectively resolves critical issues in learning behavior pattern recognition and metric weight optimization, significantly enhancing the precision and efficiency of college students' English learning engagement evaluation. This methodological advancement provides an effective analytical tool for educational big data applications, particularly in second language acquisition research..**

Keywords—machine learning, big data, data mining, automatic calculation model

I. Introduction

The sharing of learning resources through online media has been widely used in universities and colleges[1] . However, while online education is developing, it is also faces some challenges. It is important to track the students' learning engagement to predict their learning performance and to use it as an important indicator to evaluate the teaching quality and to guide the teaching management[2]. At present, scholars at home and abroad have conducted more research on student learning input methods, Liu Hong et al. proposed a metric learning algorithm incorporating Universum Learning, and Measurement Learning improves the accuracy of classification and clustering by more realistically describing the distances between samples. GMML (Geometric Mean Metric Learning) investigates the metric matrix A, which makes the same kind of distance between points as small as possible and the larger the distance between points of different classes. However, the computation time needs to be further shortened[3]. He et al. proposed to use the BPTT algorithm to analyze English error correction problem in traditional recurrent networks. Combining the BPTT algorithm with the LSTM network, they quantitatively evaluated the English speech learning and error correction. However, the accuracy of this method needs to be improved [4].

Aiming at the current problems, we construct an automatic calculation model of college students' English learning input based on big data. The experimental results show that the model effectively improves the accuracy of calculation, and through the study of students' learning behavior, it can guide students to learn independently, improve students' learning behavior, and improve students' learning performance.

II. Automatic Computational Modeling Based on Big Data

A. Data Mining Based on Big Data Technology

Before automatically calculating the amount of learning inputs of students, big data technology is used to mine the relevant data of students, and the architecture of big data is shown in Figure 1.

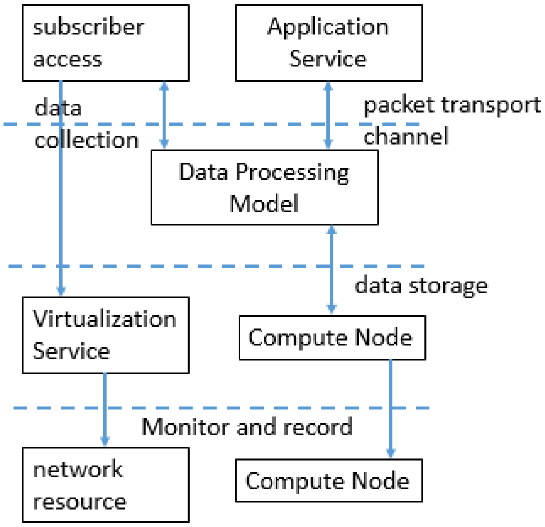

Fig. 1. Architecture of Big Data

According to the above architecture, big data technology is divided into three layers: core service layer, business management layer and user access interface layer. Data mining technology is the core technology in big data technology [5], the use of data mining technology can extract the unknown and implied information, pre-classification of data processing[6], classification formula is shown below:

$$\mathrm{S} = \frac{x}{v}/d * p \tag{1}$$

In formula (1), v represents the class label, x represents the data tuple, d represents the mapping parameter, and p represents the classification parameter.

Then there is data cleaning, that is, the processing of noisy data and isolated point data in the collected data, which is expressed as follows:

$$v_i = \frac{m/x_k}{\sum_n^{k=1}(u_{ik})^m} \tag{2}$$

In Equation (2), u_{ik} represents the fuzzy parameter, m represents the constraint parameter, and x_k represents the processing parameter of the kth indicator.

B. Online Learning Behavior Analysis Indicator Establishment

Through mining, it is found that in the process of students' online learning, there exists a large amount of data, which mainly includes structured data and unstructured data, such as achievement data, classroom data, online learning behavior records and so on[7] . And how to find valuable information from these data is an urgent problem, for this reason, in order to characterize the behavior of the learner, it is necessary to characterize the relevant attributes of the student[8], for this purpose, the relevant indicators reflecting the learning input are selected, as shown in Table 1.

TABLE I. Learning Attribute Indicators

Signal	Online Learning Behavior	Attribute Parameters
1	Students' viewing of course pages	Start and end time of viewing
2	Students' previewing of learning resources	Preview time, stop time, browsing comments, and whether there is a download action
3	Checking whether students watch course videos	Start and end time, and frequency of video viewing
4	Relevant information search	Search input
5	Downloading and saving	Number of downloads and file types
6	Received and sent file information	Sending time, sending frequency
7	Sharing of learning resources	Shared objects, time, and frequency
8	Presence of questions	Number of questions and objects
9	Course testing	Completion time and scoring
10	Self-evaluation	Quality and content of evaluation

C. Calculation of Learners' Learning Behavior Correlation

Based on the above indicator construction, the weights of the above indicators are calculated to derive the learner loyalty score and learner behavioral loyalty score [9]. The calculation formula is as follows:

$$RFL_{(study)} = \beta F' + \gamma L' + \alpha R' \tag{3}$$

In Equation (3), β, γ, and α represent the weights of F', L', and R' under the categorization index.

And analyze the characteristics of the data, in the analysis of the data is mainly to establish the correlation matrix for data analysis, correlation matrix that is the correlation matrix, assuming that (x_1 ,x_2 , x_3 ,L,X_n) is 1 n-dimensional random variable, the correlation matrix of this random variable is expressed as:

$$R = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \rho_{3n}1 & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix} \tag{4}$$

The data correlation is further calculated by evaluating and filtering the indicators to obtain meaningful correlation rules. The expression is:

$$\rho_{ij} = \frac{(x_i, x_j)}{\sqrt{DX_i}\sqrt{DX_j}} \tag{5}$$

In Equation (5), $\sqrt{DX_i}$, $\sqrt{DX_j}$ represent the correlation coefficients, and x_i, x_j represent the correlation indicators.

III. Analysis of Factors Influencing English Learning Engagement

Based on the above calculation of students' behavioral characteristics, the factors influencing learning engagement were analyzed, and the following conclusions were drawn: The Direct Role of Motivation， Motivation is the core driver of learning engagement. Research has shown that the strength of motivation has a direct impact on the degree of students' commitment to learning and is a key variable in learning engagement. Motivation motivates students to invest more energy and time in the learning process by stimulating their interest in learning and goal-oriented behavior.

A. Mediating role of learning experience

learning experience plays an important mediating role between students' learning engagement and self-directed learning. Learning experience not only reflects students' accumulation and adaptability in the learning process, but also further influences the depth and breadth of learning engagement by moderating learning strategies and methods. The accumulation of learning experience can help students better understand the learning objectives and optimize the learning process, thus enhancing learning engagement.

B. Interaction between motivation and anxiety

There is an inseparable relationship between motivation and anxiety, and their interaction affects learning engagement through the mediating variable of self-directed learning. Research has shown that students' anxiety is affected by their motivation, and changes in anxiety levels in turn affect motivation and ultimately learning engagement through the moderating effect of self-directed learning. Specifically, moderate anxiety can enhance motivation and thus promote learning engagement, while excessive anxiety may weaken motivation and lead to a decrease in learning engagement.

The learning input is calculated by combining the above behavioral indicators, the results of the correlation of the indicators and the influencing factors of learning input[10] . The model consists of the following steps:

Step 1: Divide the results of the correlation calculation of the above indicators into several groups, and then establish the cluster center and sort them;

Step 2: Normalize the sorted data and build a spatial linear transformation matrix.

$$p = \mathrm{W}\left(RF * o\left(x^{(i)}\right)\right) \tag{6}$$

In Equation (6), *RF* represents the kernel function, x(i) represents the weighted kernel function of the *ith* data, *o* represents the category volume, and *W* represents the weighted feature vector.

Step 3: Establish the student performance prediction model with the expression:

$$\mathrm{X} = \left\{X_1, \dots, X_p\right\} \in R, n * P \tag{7}$$

In Equation (7), *p* represents the feature value and *n* represents the number of learners.

Step 4: The amount of learning inputs is automatically calculated in the process shown in Figure 2 below.

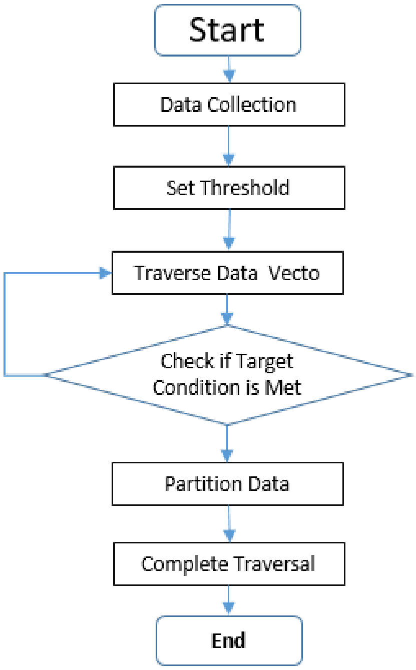

Fig. 2. Automatic Calculation Process of Learning Engagement

In the above calculation steps, the establishment of the objective function is very important and for this reason it needs to be calculated separately and expressed as:

$$\mathrm{L} = \mathrm{Y} + \sum_k \Omega(f_k) \tag{8}$$

In Equation (8), Y represents the sample size, f_k represents the decision function of the *kth* indicator, and $\sum_k \Omega$ represents the objective function.

Repeat the above process to complete the calculation of students' English learning input.

IV. Experimental Comparison

A. Experimental subjects

In order to ensure the feasibility of the designed computational model, an experimental comparison is conducted with the traditional method (literature [4]) to compare the computational effect of the two methods. Taking the online learning data of students in a certain school as an example, the data of 500 students were randomly selected as experimental data, and the 500 students were randomly divided into 10 groups, and the average of the results was taken. The main purpose of the experiment is to compare the calculation accuracy and calculation time of the traditional method and the input calculation model under study.

B. Comparison of experimental results

The results of comparing the calculation accuracy of the designed model and the traditional method are shown in Table 2 below.

The above comparison shows that the computational model has a high computational accuracy, while the traditional computational method shows unsatisfactory computational results and a low computational accuracy, with a difference of about 40% from that of the studied method. Due to the large amount of data and poor classification performance in the analysis of student behavior, the calculation effect of the traditional method is poor, which affects the calculation accuracy.

TABLE II. Comparison of Calculation Accuracy

Trial	Traditional Method (%)	Proposed Model (%)
1	86	95
2	81	92
3	80	94
4	72	92
5	74	90
6	68	98
7	50	95
9	65	91
10	62	90

Table 3 below shows the results of the comparison of the calculation time of the two calculation models. By analyzing the table below, it can be found that the calculation time increases due to the large amount of data sets analyzed, while the calculation time of the studied calculation model is less, therefore, the calculation effect of the studied method is better compared with the traditional method.

In conclusion, the automatic calculation model of college students' English learning input studied in this paper has higher accuracy and less calculation time than the traditional model. This is due to the fact that this research method integrates a variety of techniques and focuses on the use of big data technology, which makes the performance of this research method more powerful in classification and identification, and this research method can better analyze the behavioral characteristics and behavioral patterns of learners, thus improving the effect of automatic calculation of learning input.

TABLE III. COMPARISON OF CALCULATION TIME

Trial	Traditional (min)	Proposed (min)
1	15	8
2	12	8
3	14	6
4	18	7
5	16	8
6	12	2
7	13	8
9	14	8
10	14	2

V. CONCLUSION

The conclusion of this study shows that the automatic calculation model of college students' English learning input constructed based on big data technology demonstrates significant advantages in empirical evidence. By integrating learning behavior data from multiple sources, optimizing data quality by using data mining and cleaning techniques, and designing dynamic assessment indexes by combining learners' behavioral characteristics, the model achieves a double breakthrough in computational efficiency and accuracy. Experimental data show that compared with the traditional model, the accuracy of the computational model proposed in this study is improved by about 40%, while the computation time is significantly shortened, which fully verifies the rationality and technical feasibility of the method design.

The innovativeness of this study is reflected in three dimensions: first, a multidimensional data collection framework integrating big data technology is constructed, and the problem of educational data fragmentation and noise interference is effectively solved through intelligent cleaning algorithms; second, the dynamic indicator system constructed based on the characteristics of learning behaviors breaks through the limitations of the traditional static assessment, and reveals the mapping law between learning inputs and behavioral patterns through correlation analysis and state prediction model; third, the proposed calculation method has both theoretical innovation and practical value, which not only provides quantifiable analysis tools for educational researchers, but also provides data-driven decision support for the formulation of teaching intervention strategies. Third, the proposed calculation method is both theoretical innovation and practical value, which not only provides quantifiable analysis tools for educational researchers, but also provides data-driven decision support for the development of teaching intervention strategies.

Despite the expected results of the study, the following limitations still exist: first, limited by the development process of education informatization, the data collection dimension of the existing learning platforms is relatively homogeneous, and there is still a technical bottleneck in the acquisition of unstructured data, such as social interactions and affective states; second, the problem of heterogeneity of cross-platform data may affect the model's ability to generalize. Future research can be deepened in three aspects: first, developing an intelligent data acquisition system to expand the fusion application of multimodal data; second, constructing an adaptive learning analysis framework to enhance the model's adaptability to different teaching scenarios; and third, exploring feature extraction methods based on deep learning, to further enhance the depth and precision of learning law mining.

The results of this research have a positive impact on the digital transformation of education, and its technical path can provide a methodological reference for the construction of models in the field of learning analytics, as well as a theoretical basis and practical examples for universities to optimize the English teaching mode and implement precise teaching management.

ACKNOWLEDGMENT

This work was financially supported by Beijing Polytechnic University Research Fund

REFERENCES

[1] MA Yantu, SHI Qiuhong, ZHANG Zhenghuan. Research on Big Data Talent Cultivation Mode Based on Collaborative Parenting Mechanism in the Context of New Engineering+New Agricultural Science--Taking Gansu Agricultural University as an Example [J]. Journal of Nanning Normal University (Natural Science Edition), 2025, 42 (02): 7-12. DOI:10.16601/j.cnki.issn2096-7330.2025.02.002.

[2] Alatengcang. Research on Optimization and Reform of Physical Education Curriculum in Colleges and Universities Based on Big Data Analysis [J]. Sports Goods and Technology, 2025, (07): 102-104.

[3] Xue Xiaoqiang. Analysis and Research on Students' Learning Inertia in the Perspective of Educational Digital Intelligence [J]. Science and Technology Wind, 2025, (08): 140-142. DOI:10.19392/j.cnki.1671-7341.202508046.

[4] Fan Xiaohui. Exploring the development path of wisdom education in the era of big data [J]. Information System Engineering, 2025, (03): 170-172.

[5] Shao Linguang,Hao Yugang,Zhang Fei. Design of distance learning system for speech recognition based on key frame extraction and improved machine learning [J]. Automation and Instrumentation, 2025, (02): 238-242. DOI:10.14016/j.cnki.1001-9227.2025.02.238.

[6] Zhong Yuanquan. Teaching Quality Prediction in the Context of Smart Teaching-Application Based on Multimodal and Complex Networks [J]. Journal of Nanchang Engineering College, 2024, 43 (06): 82-90.

[7] Su Zizhong,Chen Yilan. Design and realization of open university webcast classroom inspection system [J]. China Informatization, 2024, (12): 49-54.

[8] Sun Bing. Teaching strategies for extracurricular reading in junior high school language under the background of "Internet+" [J]. Shanxi Education (Teaching), 2024, (12): 55-56.

[9] WANG Jiaqian,CHEN Jiajun. Application of machine learning and big data analysis in electrical equipment identification and fault warning [J]. Electronic Technology, 2024, 53 (11): 427-429.

[10] Jue Wang. Research on Mobile Learning Methods of Intermediate Computer Application Technology under the Background of Big Data [J]. Digital Communication World, 2024, (11): 232-234.