J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 2295--2310

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

An Empirical Study on Learning Path Optimization by Data Collection and Computational Analysis of English Learners' Behavior in the Era of Big Data

Xue Zhao[1], Junli Zhang[1,✉]

[1] *Teachers' College, Beijing Union University, Beijing, 100011, China*

ABSTRACT

Learning path optimization aims to generate and optimize a knowledge learning sequence for learners that best meets their knowledge needs. This study focuses on the important role of online learner behavior in personalized path planning. By constructing a knowledge point difficulty model and a learning behavior prediction model based on online learning behavior, together with a user-based collaborative filtering recommendation algorithm, a personalized learning path is proposed comprehensively. The MOOC websites "College English 1" and "Xuedang Online" are selected as sample data to analyze the online learning behavior of English learners and verify the learning effect of the learning path proposed in the article through the change of students' online time. The personalized teaching model based on the learning path is investigated in practice by taking the college English course in school A as an example. Compared with the traditional teaching mode, the optimized learning path shows a significant difference of 0.01% in the dimensions of learners' "knowledge and skills", "process and method" and "affective attitude". The mean values of the optimized blended teaching mode are 4.12, 4.33 and 4.07 respectively, which are all better than the traditional teaching mode. It shows that the English learning path proposed in this paper is conducive to enhancing students' personalized learning needs and provides a reference for promoting the effective implementation of personalized learning in the information technology environment.

Keywords: knowledge point difficulty model, learning behavior prediction, collaborative filtering recommendation algorithm, learning paths

1. Introduction

Nowadays, college students' English learning is no longer simply mastering basic language skills, but involves the cultivation of interdisciplinary knowledge, cross-cultural communication and comprehensive ability. Understanding and analyzing the data of college students' English learning

✉ Corresponding author.
E-mail address: changcarla@126.com (J. Zhang).
Received 10 January 2024; Revised 25 May 2024; Accepted 20 October 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-130

behaviors can provide a reference for realizing personalized teaching and improving students' learning effects [1-2]. The data on college students' English learning time, the use of learning resources, and the choice of learning styles not only reflect the dynamics of students' learning, but also provide a strong basis for improving educational methods [3]. College students have different motivations and learning habits, and these learning behavior data need to be analyzed in depth in order to develop more targeted educational programs [4].

Learning behavior measurement has become one of the trends in the development of educational technology [5]. Many studies have also shown that learning behavior input and academic performance are inextricably linked, and how to use modern information technology to scientifically quantify and explain the relationship between the two is particularly important, and thus to carry out research on the measurement and prediction of learning behavior input in college English, and to realize the in-depth integration of information technology and classroom teaching, is an important topic for college English teachers to carry out teaching research [6-7].

Big data is a new product of the development of Internet technology, which contains cloud computing, distributed processing, storage and perception technologies [8]. It makes the original not easy to find, difficult to collect, impossible to use the massive variety of data generated in the process of production and life, began to enter the field of vision of the researchers, easy to collect and categorize, can be analyzed to enhance, so for the progress of human society and development of the value created by the exponential growth [9-10]. Compared with the data obtained by traditional means, big data has a huge amount of information, unstructured distribution, strong accessibility, and intelligent analysis, directly facing the ordinary users, everyone can be a data user, breaking the shackles of only professional and technical personnel can get involved [11-12].

In traditional classroom teaching, the teacher's judgment of students and the organization of teaching behavior is based on a short period of time, small scenes, small events between the results and the possible causes of the relationship between the results, more by virtue of experience to sort out, identify and explore, the accumulated experience is weak scalability [13-14]. At the same time, teachers are unable to depict the process of quantitative to qualitative changes in learning or teaching activities because they cannot access more specific and continuous data information, and tend to derive the relationship of influence and the direction of change from the meso level, with limited means of accessing data on teaching and learning activities at the micro level, and limited ability to analyze data at the macro level [15-16]. In contrast, university English teaching in the big data environment will no longer rely solely on concepts and experiences, but on empirical teaching with data and rationale. Big data technology can be applied to develop individual and group learning programs for English learners, analyze their learning behaviors and tendencies, guide and monitor the learning process, identify weak links, predict learning needs, and enrich the learning evaluation methods [17].

In the research of big data-enabled English teaching, some scholars focus on big data technology to facilitate the construction of English special education and personalized teaching environments. Qian, R et al. envisioned the use of Big Data Integrated Artificial Intelligence with AAC (BDIAI-AAC), which in turn assisted the English learning of neurologically challenged patients, and in practice, this method achieved a 98.01% recognition rate of English vocabulary [18]. Sun, M et al. who conceptualized a new ELT ecosystem building strategy that contributes to the construction of a personalized, high-quality ELT ecosystem [19].

Some scholars have also explored the role played by big data technology around the perspective of English teaching evaluation and students' behavior analysis. Jing, Y et al. proposed an efficient English teaching evaluation model with big data analysis technology as the core logic, and verified the effectiveness of the assessment model by means of a controlled experimental class [20]. Tan, Q built a framework for English teaching evaluation based on big data technology, and revealed that the teaching of English teachers who are in the age of 45-55 is more popular among students [21]. Wang, Q et al. tried to introduce computerized big data technology in the reform of English teaching and optimized the model of this teaching reform by using fuzzy fusion processing method, which realized the accurate evaluation of English teaching indexes [22]. Wang, C developed a data mining technology-based logical online English teaching student behavior analysis method, which shows the advantages of high efficiency and small prediction error in practical application [23].

In this study, a knowledge difficulty score model based on learner behavior was constructed from the perspective of the implicit relationship that exists between learners' online learning behavior and knowledge point difficulty. The weights of the knowledge difficulty model were generated using hierarchical analysis. And a self-attention mechanism with time and location signals is introduced in the generalized hypergraph neural network framework (HGNN) for representing the constructed directed event hypergraph, and the node features are represented using the in/out degree weight matrix of the graph to predict the learning behavior. On this basis, a user-based collaborative filtering algorithm is utilized to give learning element sequence recommendations to enrich the learning element recommendation list, and the recommendation model will match different recommendation strategies for different learners according to the learner portrait and the actual learning progress and state of the learners.

2. Model construction and learning path recommendation design

2.1. Construction of Knowledge Point Difficulty Model Based on English Learning Behavior

Online learners of English will repeatedly watch clips that they find difficult, and learners' test scores also reflect to some extent how difficult a particular question is for the individual learner. Second, when learners encounter bottlenecks that are difficult to break through in the process of English learning, they also tend to post or reply to comments in the comment section to ask for help or discuss. However, through a literature search of previous studies, the author found that the details of these learners' English learning behaviors were overlooked in the question of judging the difficulty of knowledge points. Therefore, from the perspective of the implicit relationship between learners' online learning behaviors and the difficulty of knowledge points, which has been neglected in the past, this study takes into account the relationship between a series of online learning behaviors (including video viewing behaviors, forum interaction behaviors, and practicing behaviors) and the difficulty of knowledge points when trying to judge the difficulty of knowledge points. In this study, in order to abstract the relationship between different parameters, the author constructed a model for measuring the degree of difficulty of a specific knowledge point for the general students, i.e., a knowledge point difficulty score model based on learners' behaviors. The specific formula is shown in Equation (1) under Fig.

$$diff_{(j)} = \omega_1\left[1-\overline{sco}(j)+\omega_2\overline{rep}(j)+\omega_3\overline{com}(j)\right] \tag{1}$$

Where $\overline{rep}(j)$ and $\overline{com}(j)$ represent the quantitative results of the normalized data of learners' learning behaviors during the English online learning process. $\overline{sco}(j)$ is the practice test score obtained from the online learners' practice after watching the video.

In Eq. (1), ω_1, ω_2, and ω_3 are the weights of the input parameters, $\overline{sco}$, $\overline{rep}$, and $\overline{com}$ are the average test scores of the i th knowledge point for the j th student, the total number of repeated viewings of the video of the i th knowledge point for the j th student, and the total number of discussions of the i th knowledge point for the j th student, respectively. The details are calculated as in equations (2), (3) and (4).

$$\overline{sco}(j) = \frac{\sum_{i=1}^{N_j^{Aistory}} scoij}{N_j^{Aistory}} \tag{2}$$

$$\overline{rep}(j) = \frac{\sum_{i=1}^{N_j^{Aistory}} repij}{N_j^{Aistory}} \tag{3}$$

$$\overline{com}(j) = \frac{\sum_{i=1}^{N_j^{Aistory}} comij}{N_j^{N_j^{Aistory}}} \tag{4}$$

The input parameter of the knowledge difficulty model is the average historical learning performance $\left(1-\overline{SCO}(j)+\omega_2\overline{rep}(j)+\omega_3\overline{Com}(j)\right)$ of all users who have learned the knowledge point, and the output is the difficulty of mastering the knowledge point $\left(diff_{(j)}\right)$. After the calculation, a larger value of $dif\ f_{(j)}$ means that the knowledge point is more difficult for the student.

ω represents the weight of the knowledge difficulty model generated by the hierarchical analysis method. The calculation of ω is mainly performed by means of hierarchical analysis (AHP). Hierarchical analysis is a multi-criteria decision-making method that is simple, flexible and scientifically practical for quantitative analysis of qualitative problems. As an adaptive positive hypothesis testing method, hierarchical analysis has the advantages of controlling the error rate, high efficiency, wide applicability, wide theoretical basis and application fields, and can be applied to hypothesis testing problems in various fields. Hierarchical analysis method is characterized by the various factors in the complex problem by dividing them into interconnected and orderly levels to make them organized, according to the structure of subjective judgment of a certain objective reality to combine the expert's opinion and the results of the analyst's objective judgment directly and effectively, and quantitatively describing the importance of two-by-two comparisons of the parameters at one level. After that, the mathematical method is used to calculate the weights reflecting the order of relative importance of parameters at each level, and the relative weights of all parameters are calculated and ranked by the total ranking between all levels.

After calculation, the CR value is 0.028, which is less than 0.1. As a result, the results passed the consistency test.

2.2. *Learning Behavior Prediction Model*

The learning behavior prediction problem studied in this paper can be described as follows: during the learning process of a course, given the set of historical learning events S^u of the target learner for the course and the historical learning events S of all learners, predict the course resources r that the target learner u will learn next. Construct a probability function f to compute the likelihood $\hat{y}$ of a learner u interacting with a course resource r that is not undergoing learning in the event hypergraph G_H consisting of all learners' learning events, defined as:

$$\hat{y} = f\left(u, r \mid S^u, S; \theta\right) \tag{5}$$

where Θ denotes the argument of function f.

2.2.1. Event hypergraph embedding. To determine the parameter θ of function f in Eq. (5), this paper introduces a self-attention mechanism with time and location signals in the generalized Hypergraph Neural Network framework (HGNN) to capture the dynamic learning preferences behind learning events by considering the learning time and different levels of attention to learning resources in the learner's historical behavioral data.

In the model of this paper, the nodes in the hypergraph are directed, which is ignored by the traditional HGNN framework when embedding nodes. Therefore, in this paper, in order to represent the constructed directed event hypergraph, the node features are represented using the in/out degree weight matrix of the graph. Assuming that for n node $s_t = \{r_1^t, r_2^t, \ldots, r_n^t\}$ contained in each learning event in the hypergraph, their feature matrices are denoted as $V^c = \left[v_1^c, v_2^c, \ldots, v_n^c\right]^T$, their input information x_i can be represented as:

$$x_i = \left(D_{s,i}^{out} V^c \oplus D_{s,i}^{in} V^c\right) W + b \tag{6}$$

where $D_{s,i}^{out}$, $D_{s,i}^{in} \in \mathbb{R}^{1\times n}$, represent row i of the corresponding matrix, and $W \in \mathbb{R}^{2d\times d}$ and $b \in \mathbb{R}^{1\times d}$ are two learnable model parameters that can be derived without difficulty $x_i \in \mathbb{R}^{1\times d}$.

Based on the above initial feature representation of the directed hypergraph nodes, the initial embedding of the hypergraph nodes is $X^0 = \left[x_1^0, x_2^0, \ldots, x_{|V|}^0\right]$ and the convolution layer can be defined as:

$$x_i^L = \sigma\left(\sum_{v=1}^{|V|}\sum_{e=1}^{|E|} H_{ie} H_{ve} W_H x_v^{(L-1)} Q^{(L-1)}\right) \tag{7}$$

where $\sigma(\cdot)$ represents the nonlinear activation function (ReLU) and $Q^{(L-1)}$ represents the trainable weight matrix of layer $L-1$. This convolution operation encodes each hyperedge with all the nodes it is connected to, and then outputs the embedding of each node by aggregating the information of all the hyperedges it is on. The above convolution process can be written in matrix form, and symmetric normalization needs to be added in order to prevent numerical instability caused by stacking multiple convolutional layers:

$$X_t^L = \sigma\left(D_v^{-\frac{1}{2}} H W_H D_e^{-1} H^T D_v^{-\frac{1}{2}} X_t^{(L-1)} Q^{(L-1)}\right) \tag{8}$$

where X_t^L and $X_t^{(L-1)}$ denote the hypergraph embedding vectors of layers L and $L-1$, W_H is the diagonal matrix of hyperedge weights, and $Q^{(L-1)}$ is the trainable weight matrix of layer $L-1$, which is continuously updated by training the model to finally extract the node features of the hypergraph.

By averaging the node embeddings obtained from the L-layer convolution, the embedding representation of the nodes can be finally obtained:

$$X_t = \frac{1}{L+1}\sum_{l=0}^{L} X_t^l \tag{9}$$

Based on the above learning training process, the embedding matrix $Z = \left[Z_1, Z_2, \ldots, Z_{|E|}\right]$ of the nodes in the learning event hypergraph of learner u can be obtained, where $Z_t = \left[x_1^t, x_2^t, \ldots, x_n^t\right]$ denotes the feature representation of the nodes in each hyperedge.

2.2.2. Prediction of learning behavior. With the representation of learners' dynamic learning preferences based on the temporal attention mechanism, a corresponding score can be computed for each candidate learning resource r_i in the event hypergraph that the learner does not interact with, as represented below:

$$\hat{z}_i = I^T x_i \tag{10}$$

The probability of each course resource being the next item was then calculated using the softmax function:

$$\hat{y} = soft\max(\hat{z}) \tag{11}$$

Finally, the cross-entropy loss function is utilized for prediction:

$$\mathrm{L}_r = -\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right) \tag{12}$$

2.3. User-based collaborative filtering to recommend learning paths

On the basis of the existing learner group segmentation, in order to fully reflect the learner's subjectivity, this study tries to utilize the user-based collaborative filtering algorithm to give learning meta-sequence recommendations to enrich the learning meta-recommendation list.

The following is the formula used in the learning path recommendation

$$I = equal\left(S_0^C, S_X^C\right) = \begin{cases} 1 & if\ S_0^C = S_X^C \\ 0 & if\ S_0^C \neq S_X^C \end{cases} \tag{13}$$

S_0 and $S_x\left(x = 1, 2, \cdots, n\right)$ denote the user learner and the candidate learner respectively, S_0^C and S_X^C denote the learning content of the learning element in the user learner and the candidate learner, and $equal\left(S_0^C, S_X^C\right)$ is a binary function that determines whether S_0 and S_x are the same, and if they are the same, then it returns 1, otherwise it returns 0. The similarity and the average fluency are calculated as follows:

$$sim\left(S_0, S_x\right) = \frac{count\left(S_0, S_x\right)}{len\left(S_0\right)} \tag{14}$$

$$mf(S) = \frac{1}{len(s)} \sum_{e \in S}^{len(s)} f(e) \tag{15}$$

$sim\left(S_0, S_x\right)$ denotes the similarity between user learner S_0 and candidate learner S_x, $count\left(S_0, S_x\right)$ denotes the number of identical learning elements in the lists S_0 and S_x, and $\text{len}(s)$ denotes the length of the learning path . $mf(S)$ denotes the average correct fluency of learner S. E is the learning element included in the path of learner S, and $f(e)$ is calculated as the correct fluency of learning element e.

$$p\left(S_X\right) = \frac{sim\left(S_0, S_X\right) + mf\left(S_X\right)}{T^t} + e^t \tag{16}$$

The $p\left(S_X\right)$ value can be used as an indicator for ranking candidate learners. T^t indicates the time spent switching between two learning elements, and e^t indicates the time spent by the learner in the learning process of learning element e. Both can be obtained directly from the learning activities of the learning element. $p\left(S_X\right)$ The similarity between the candidate learner and the user learner as well as the average correct fluency of the candidate learner are also considered, and the time spent on switching between learning elements and the learning duration of learning elements are also taken into account. When the similarity between candidate learners and user learners as well as the average correct fluency of candidate learners are higher, the time spent on learning element switching is lower, and the learning element learning duration is higher, $p\left(S_X\right)$ the value is higher.

Based on the details of the learner's portrait in the original database of learner portraits and the learner's actual learning progress and status, the recommendation model will match different recommendation strategies for different learners. For the A types of target learners, the most prominent feature is that there is no or only a small amount of learning history data available, and recommendations are given with the help of the default learning meta-sequence and the learning meta-sequence of the excellent learners given in the instructional design, and the former is prioritized.

For category B learners, since they have completed one or more courses, there is a certain amount of learning data in the e-learning platform, and with the existing learner profile tags, we find the learner group corresponding to this learner Q. In addition, according to the target learning content S_0^C, we search for learners in the learning meta-sequence with the same content S_X^C to form a list of candidates, and sort the list of candidate learners by the average correct fluency, from high to low. The list of candidate learners is sorted by the average correct fluency in descending order, and the top n

candidate learners are selected. The learning element sequences of these n learners, together with the learning element sequences of the excellent learners in the above Q and the default learning element sequences in the instructional design, are pushed to the learners, and the learners will choose one of the learning elements for learning independently.

For type C learners, the learning data of the target course has been left in the platform, firstly, the learner group division is completed by the learner portrait to obtain the learner group Q_1 including the learner, and then the learning content S_0^C of the last learning element of the user's learner's historical learning data is used as a match, and the S_X^C with the same learning content are found to form the candidate learner group Q_2, and then the average correct fluency of the learning elements learned by each learner in Q_1 and Q_2 is used $mf(S)$. Sort the list of candidate learners in descending order as a condition; Finally, the top n candidate learners are selected, and the learning meta-sequences of these n learners in the learning target unit are recommended to the target learners, together with the default learning meta-sequences given by the instructional designer, and the learners choose one of them independently.

For the target learner of type D, who already has a large amount of historical learning data, he already has a more stable learner profile and learner groups. Similar to the target learners of category C, the learner group Q_1 is obtained by dividing the learner profiles into groups. Taking all the learning elements they have learned as the matching items, and with the threshold value of t as the condition, find the learners whose similarity and average fluency are greater than t to form the candidate learner group Q_2; and then arrange the candidate learner groups Q_1 and Q_2 in order of $p\left(S_X\right)$ from the highest to the lowest, and then take the learning element sequences of the first n candidate learners when they learn the target unit as well as the default learning element sequences to be recommended to the target learner, and then the learner will choose one of them independently. Learning.

3. Empirical study of learner behavior data analysis for learning path optimization

3.1. *Analysis of English Learners' Online Learning Behavior*

3.1.1. Sample Selection and Data Processing. In this study, first-year students from a university in Guangdong Province were selected as the observation point, and the data provided by the "Xuetang Online" participating in the MOOC "College English 1" course in the fall semester of 2023 (from September 1, 2023 to February 17, 2024) were collected, and the data analysis and mining such as descriptive statistics, correlation analysis, analysis of variance, and cluster analysis were carried out by IBM's statistical product and service solution SPSS28.0, supplemented by content mining software ROST-CM6 for content analysis and mining. This paper describes, counts and analyzes the learning behaviors of learners in public English courses on the new learning system, reveals the group characteristics of learning behaviors, excavates the reasons behind them, and provides reference for the current open education, teaching, learning, assessment and management.

According to the "Xuetang Online" data framework provided by the MOOC website "College English 1", an analysis model of 8 indicators of learners' online learning behavior was constructed, including: number of visits, visit duration, audio and video viewing rate, online link viewing rate, online test submission rate, homework submission rate, discussion participation rate, and course completion rate. In addition, content analysis is performed on Q&A forum posts.

First, Microsoft Excel was used to preliminarily process the collected "learning analysis" data, and then SPSS28.0 was used for further data statistics and analysis. The results show that from September 1, 2023 to February 17, 2024, a total of 5,721 learners registered in the university's "College English 1" online course.

3.1.2. Online learning behavior. The descriptive statistics of online learning behaviors are shown in Table 1. Excluding the 106 learners who have never visited, the statistical results of the 5,615 recorded learners suggest that the time invested in learning is relatively small, the degree of completion of learning is still acceptable, the utilization rate of learning resources is not high, there is not much interaction, and the depth of learning is insufficient. Taking an average of 3 months and 90 days as a learning cycle, the average number of visits per day is less than 3, and the average length of a visit is only 1 minute. The rate of audio and video viewing was 37.73%, the rate of online link viewing was 13.43%, the rate of homework submission was 61.64%, the rate of online quiz submission was 78.48%, the rate of participation in discussion was 16.29%, the course completion rate was 81.64%, and the average score of the final online test was 61.83. In addition, the statistical results also show that, except for the online quiz submission rate, the variance values on the other dimensions are very large, indicating that the learning behaviors among different learners are very discrete, and that there is a serious differentiation between highly motivated learners and lowly motivated learners. Whereas the online quiz was the only module that counted towards the formative assessment, learners were not too differentiated here.

Table 1. Descriptive statistics of online learning behavior(N=5721)

	Mean value	Variance	Maximum	Minimum
Number of visits X1	253	37392	14823	0
Access duration minutes X2	263.93	63936.54	1398.48	0.37
Audio and video viewing rate X3	37.73%	20.71%	100.00%	0.00%
Online link view rate X4	13.43%	15.29%	100.00%	0.00%
Homework submission rate X5	61.64%	12.78%	100.00%	0.00%
Test submission rates online X6	78.48%	5.23%	100.00%	0.00%
Discussion participation rate X7	16.29%	18.83%	100.00%	0.00%
Course completion degree X8	81.64%	9.73%	100.00%	0.00%
Final grade X9	61.83	20.65	100	0

3.1.3. Analysis Related to Online Learning Behavior. Correlation analysis using Pearson's correlation analysis is shown in Table 2, and the results show that there is a significant positive correlation between all behaviors, except for the number of visits and length of visits, which do not have a significant correlation with the assignment submission rate. The highest correlation coefficient between course completion and discussion participation was 0.886, indicating that students with high discussion participation also performed well in course completion.

Table 2. Correlation analysis of online learning behavior

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1								
X2	0.582**	1							
X3	0.666**	0.557**	1						

X4	0.583**	0.548**	0.726**	1					
X5	0.203	0.493	0.589**	0.286**	1				
X6	0.382**	0.442**	0.367**	0.640**	0.403**	1			
X7	0.508**	0.579**	0.316**	0.342**	0.686**	0.458**	1		
X8	0.293**	0.604**	0.467**	0.743**	0.666**	0.720**	0.886**	1	
X9	0.579**	0.736**	0.821**	0.646**	0.650**	0.755**	0.533**	0.828**	1

** indicates significant correlation, $p < 0.05$

3.1.4. Cluster analysis. After Z-score standardization of the data, cluster analysis of the raw data using K-means clustering is shown in Table 3, which shows that there are 673 active learners, accounting for 11.76%, 1,684 relatively active learners, accounting for 29.44%, 2,463 generally active learners, accounting for 43.05%, 789 learners who are not active in learning, accounting for 13.79%, and no learners 112, or 1.96%.

Table 3. Cluster analysis results

	1	2	3	4	5
Number of visits	804	515	249	155	3
Access duration minutes	1038.38	693.43	284.23	148.32	2.03
Audio and video viewing rate	0.904	0.753	0.535	0.402	0.011
Online link view rate	0.893	0.686	0.337	0.179	0.008
Homework submission rate	0.952	0.708	0.605	0.482	0.014
Test submission rates online	0.989	0.841	0.783	0.487	0.035
Discussion participation rate	0.883	0.680	0.207	0.103	0.000
Course completion degree	1.000	0.904	0.829	0.638	0.007
Final grade	94.83	80.32	60.23	40.62	20.38
Number of people	673	1684	2463	789	112
Take up a proportion of	11.76%	29.44%	43.05%	13.79%	1.96%

3.2. Learning Effectiveness Test

Based on the previous data collection and computational analysis of English learners' behaviors in the era of big data, the article proposes to provide learners with adaptive learning resources and learning optimization paths, in order to improve learners' academic performance and enhance learning effects. This study verifies the learning effect through the change of students' online time.

The change of online time reflects the learners' preference and dependence on the optimized learning path to a certain extent. Due to the large amount of data, when studying the change of online time, several students can be selected to count the weekly online learning time in the range of 3 months. Each student's weekly study hours are made into a two-dimensional table and a bar graph is drawn to observe the weekly online time changes. Figure 1 shows the effect of the test data in the system, randomly selected 10 user IDs (S1-S10) 12 weeks of online time statistics.

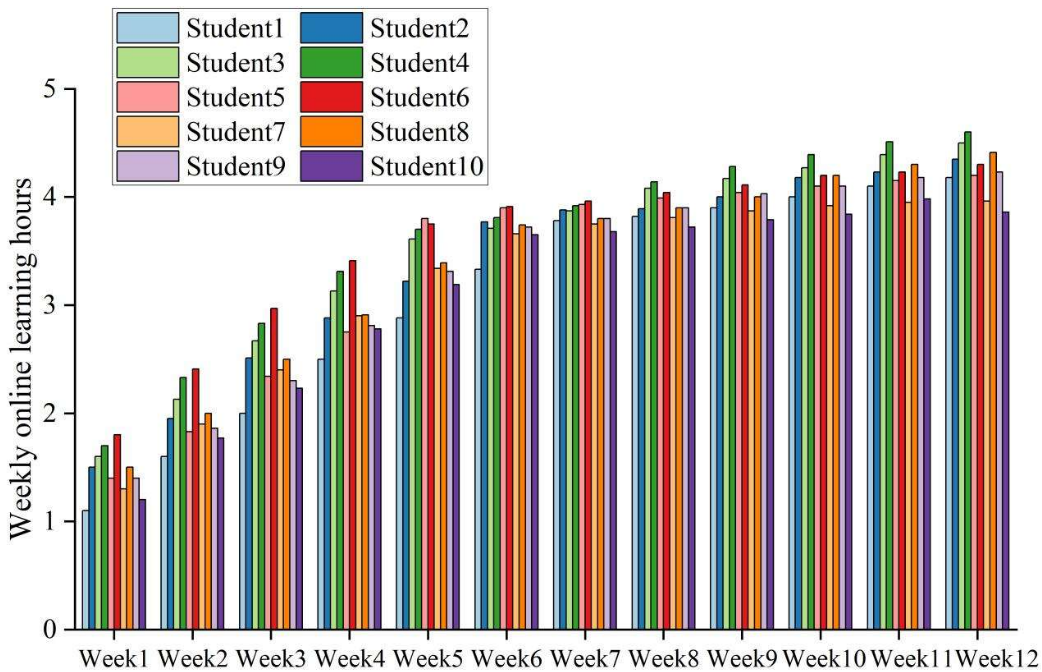

Fig. 1. Online learning time statistics

From Figure 1, it can be seen that the average weekly online learning time of students in the first week is 1.45 h, and it gradually improves every week, and stabilizes in the sixth week, improving to 3.87 h. If the online learning time is shorter at the beginning, after a period of time, the general growth of the online learning time may be that they are not too familiar and not too adapted to the system. From the background learning data, we can find out how much time students spend every week, in which time period they usually enter the system to learn, and whether they have a fixed time to enter the system to learn. Explore whether the time students log into the system to study is related to the time they attend offline English courses, and whether students are accustomed to entering the system before and after classes to prepare for or review after classes. The overall online learning time tends to stabilize once it reaches a value, indicating that to a certain extent students have become dependent on the system, suggesting that the optimized learning path is still more helpful for learning.

3.3. Survey on the Effectiveness of English Teaching under Optimized Learning Paths

In order to better test the effectiveness of personalized learning based on the collaborative filtering recommended learning path, we find out whether the optimized personalized learning path can effectively meet the differentiated learning needs of learners and promote the achievement of the goal of learners' "learning to learn", so as to ultimately promote the realization of personalized learning. This study investigates the personalized teaching mode based on collaborative filtering recommended learning paths in a college English course in Guangdong Province (hereinafter collectively referred to as "School A"), in order to better grasp the impact of optimized personalized teaching on learners' personalized learning.

3.3.1. Subjects and Methods. In order to investigate the teaching effectiveness of user-based collaborative filtering recommended learning paths, this study takes the teaching mode as the independent variable and the optimized personalized teaching mode as the dependent variable.

The study takes the teaching mode as the independent variable, which mainly includes the traditional teaching mode and the optimized personalized teaching mode, and tries to control the interfering variables such as teachers, teaching content, students' level, teaching time, etc., and observes the dependent variable of students' learning effect, that is, the achievement of personalized learning objectives, in order to investigate the impact of the personalized learning mode of collaborative filtering and recommended learning paths on the development of students' autonomy.

This study selected School A as the research object, which has complete hardware facilities, significant achievements in education informatization, and mature conditions for blended teaching. The research subjects of this study are the students of four classes in the College of Business Administration in the first year of university A, and the total number of samples is 276. In the experiment, the average of the students' scores on the university English exam in each class was used as the reference quantity, and the overall level of the students in each class was similar, and the distribution of the students' levels within the classes was also similar. In addition, the teaching experience and style of the instructors of the English course are similar, and the teaching content and progress are the same, which can effectively exclude the interference of other irrelevant variables.

Before the beginning of the experiment, a pre-test was conducted for the participating classes, on the one hand, in order to understand the learning situation of the students; on the other hand, the pre-test is also an important reference for the analysis of the subsequent study. A post-test was to be conducted after the experiment to investigate the students' learning effects and the achievement of their individualized learning goals against the students' learning effects under different teaching modes.

3.3.2. Comparison of Personalized Learning Effectiveness in Optimized Pathways and Traditional Instructional Environments. In this study, the experimental method was used to collect data using questionnaires on the teaching situation of the experimental group that carried out user-based collaborative filtering to recommend learning paths for personalized learning instruction and the control group that was taught traditionally, and SPSS 20.0 was used to analyze the data.

In order to test the difference between the effects of personalized learning in the optimized blended teaching environment and the traditional teaching environment, this study analyzed the differences in the achievement of personalized learning goals between the experimental group and the control group. Table 4 shows the difference analysis of personalized learning effect in the two modes.

Table 4. The difference analysis of personalized learning effect in different modes

Dimensionality	Class	Sample size	Mean value	Standard deviation	T	P
Knowledge and skills	Traditional teaching class	137	3.34	0.6943	0.051	0.00
	Optimized teaching class	139	4.12	0.6834		
Process and method	Traditional teaching class	137	3.58	0.7403	0.046	0.00
	Optimized teaching class	139	4.33	0.6393		
Emotional attitude	Traditional teaching class	137	3.25	0.6820	0.040	0.00
	Optimized teaching class	139	4.07	0.7035		

A t-test was applied to examine the differences in the dimensions of personalized learning objectives between two different instructional models, namely, blended instruction and traditional instruction, based on users' collaborative filtering of recommended learning paths. The results showed a significant difference of 0.01 in the dimensions of learners' "knowledge and skills", "process and methods" and "emotional attitudes" of different modes of teaching activities (P=0.00<0.01), and by comparing the mean values of 4.12>3.34, 4.33>3.58, 4.07>3.25, it was found that the average value of the optimized mixed teaching class was greater than that of the traditional teaching class. The blended teaching model based on user-based collaborative filtering and recommended learning paths can promote personalized learning for students. This is due to the user-based collaborative filtering recommended learning path of personalized teaching mode from the problem situation, guiding students to carry out independent, cooperative inquiry learning, and focusing on the independent development of students in the learning process, which can stimulate the active emotional participation of students and better contribute to the achievement of teaching goals.

3.3.3. Correlation analysis between personalized learning needs and personalized learning objectives:

1) Satisfaction of personalized content needs is conducive to personalized learning objectives. The learners' content needs and personalized learning objectives are tested for correlation analysis, and the Pearson correlation coefficient is used to indicate the strength of the correlation. First of all, the P-value of each dimension of content needs and the overall personalized learning objectives, which is also the Sig value in the table, is less than 0.01, which proves that the learners' content needs and personalized learning objectives show significance at the 0.01 level. And the R-values are all positive, indicating a positive correlation between the two. The specific correlation coefficient values of the dimensions of content needs and the dimensions of personalized learning objectives are shown in Figure 2, and the R-values are all in the range of 0.5 to 0.8, indicating that there is a close correlation between learning activities and personalized learning objectives.

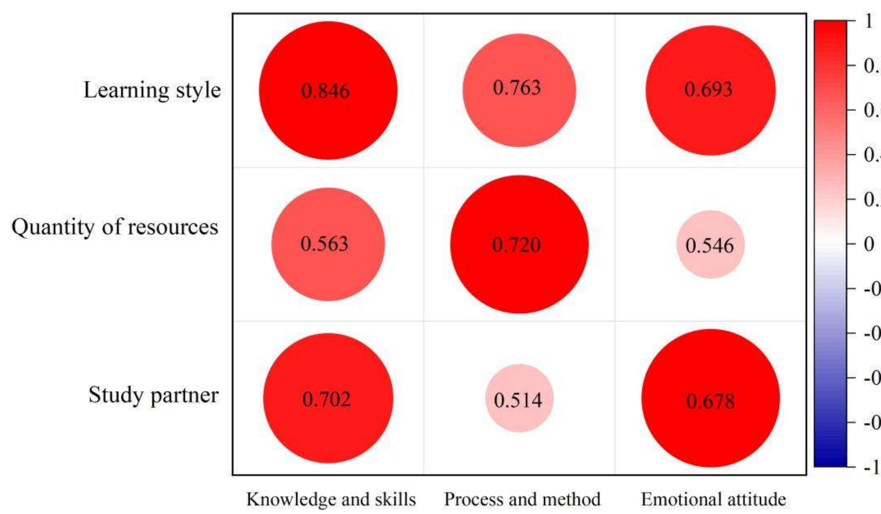

Fig. 2. Correlation analysis between content demand and learning goal

The blended teaching mode of recommending learning paths based on collaborative filtering of users can not only make full use of the network to provide learners with learning content and learning resources in different forms, at different levels, and at different levels of difficulty, but also set up and push teaching content as well as learning tasks and assignments suitable for the level of students at different levels according to the level basis of the learner, which is conducive to improving the learners' self-efficacy, confidence in learning and more interest in learning by completing learning tasks suitable for their own levels. By completing learning tasks at their own level, learners' self-efficacy can be improved, and learners can be more confident and interested in learning. Especially in a language subject like English, increased self-efficacy helps students to express themselves bravely and improve their ability to express themselves and do things in English. Moreover, the personalized learning path allows learners to choose a variety of learning contents according to their own preferences and levels to a certain extent, which greatly enhances students' interest in learning, and then promotes the change of their learning attitudes, the enhancement of their intrinsic motivation, and the improvement of their will to learn.

2) The satisfaction of personalized process needs is conducive to the realization of personalized learning goals. The correlation analysis between the dimensions of learners' process needs in learning and personalized learning goals shows that there is a high correlation between the satisfaction of learners' process needs and the achievement of personalized learning goals, and the correlation analysis between process needs and personalized learning goals is shown in Figure 3. First of all, the P-value (Sig value) of learning progress, interaction mode and interaction frequency dimensions under the dimension of learners' process needs and the dimensions of personalized learning goals = 0.000<0.01, which proves that the process needs of learners and personalized learning goals show significance at the 0.01 level. And the correlation coefficients R-values are all positive, indicating a significant positive correlation between the two. The correlation coefficients of learning progress, interaction mode, and interaction frequency dimensions with the dimensions of personalized learning goals are all greater than 0.6, indicating that the process needs have a close influence on the personalized learning goals, especially on the learners' affective attitudes, which show high correlation.

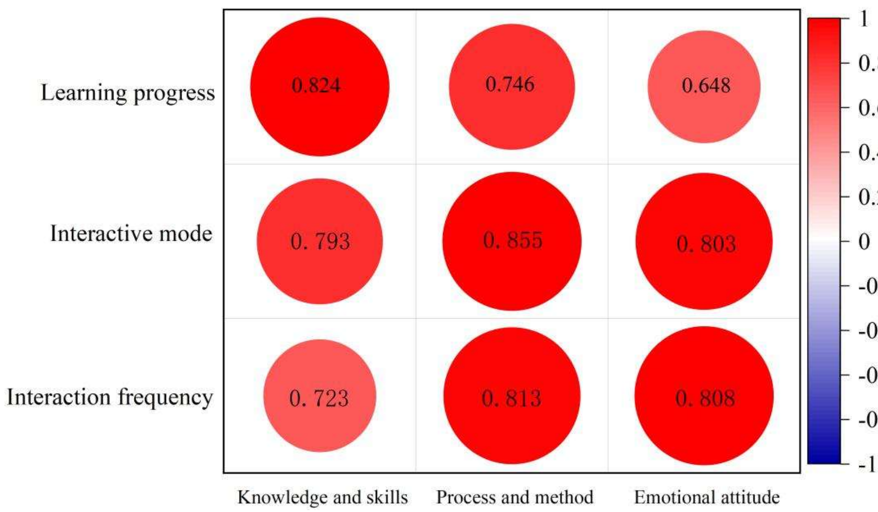

Fig. 3. Correlation analysis between process requirement and learning goal

Through data analysis, it is found that the blended teaching mode based on the user's collaborative filtering and recommended learning paths can well meet the learners' needs for personalized learning process in terms of learning progress, interaction mode and interaction frequency. Blended teaching effectively combines online and offline teaching, and students can conduct independent inquiry learning based on the resources pushed by the teacher, which promotes the formation of students' "personalized knowledge". Learners also have greater autonomy over their learning time and progress, and they can make full use of the cognitive tools provided by the platform, such as mind maps and knowledge maps, to better master the learning method. In the blended teaching environment, teachers also pay more attention to the cultivation of students' learning styles, guiding learning to learn independently, cooperatively, and through inquiry.

In addition, the use of online platforms for blended teaching provides more real-time interaction opportunities for learners, improves the shortcomings of traditional teaching in terms of interaction, enriches the interaction mode, increases the frequency of interaction, and expands the interaction surface. On the one hand, through the online teaching platform, learners can instantly communicate with the teacher online, and the teacher can also provide timely help for students, improving the traditional teacher-student interaction, allowing learners to have the opportunity to express their inner feelings, and greatly improving the enthusiasm of learners' participation in the classroom.

4. Conclusion

The article proposes a personalized learning path based on users' collaborative filtering recommendation by constructing a knowledge point difficulty model and a learning behavior prediction model based on English learning behavior, collects and computationally analyzes English learners' behavioral data, and conducts a practical research on the optimized personalized teaching path, with the following conclusions.
1) The homework submission rate of online learning behavior in college English courses is 61.64%, the online quiz submission rate is 78.48%, the course completion rate is 81.64%, and the mean value of final online test scores is 61.83, which indicates that the students' performance in online learning is relatively moderate.
2) Cluster analysis divides learners into 5 categories according to learning behavior, with active learners accounting for 11.76%; relatively active learners accounting for 29.44%; generally active learners accounting for 43.05%; learners who are not active learners accounting for 13.79%, and no learners accounting for 1.96%.
3) The learning effect test shows that the average weekly online learning time of students is 1.45h in the first week, which is raised to 4.28h in the twelfth week.
4) The average values of learners in the three dimensions of "knowledge and skills", "process and methods" and "emotional attitude" in the optimized mixed teaching class were 4.12, 4.33 and 4.07, respectively, which were larger than those of the traditional teaching class.
5) The specific correlation coefficients between the dimensions of content needs and process needs and the dimensions of personalized learning goals are in the range of 0.5 to 0.8, indicating that there is a close influence between the two demand dimensions and the personalized learning goals.

The study fully demonstrates that the optimized English learning path proposed in the article is effective and provides new ideas to promote the solution of personalized learning problems.

References

[1] Wang, Q. (2021, November). Application of recommendation algorithm and big data technology in computer english corpus database. In Journal of Physics: Conference Series (Vol. 2083, No. 3, p. 032092). IOP Publishing.

[2] Fang, M. (2023). Forestry english corpus construction and application in foreign language teaching under the background of big data and internet of things. Journal of computational methods in sciences and engineering, 23(1), 211-221.

[3] Yu, L., & Shen, J. (2022). Analysis of the Correlation between Academic Performance and Learning Motivation in English Course under a Corpus - Data - Driven Blended Teaching Model. Scientific Programming, 2022(1), 3407270.

[4] Wei, X., Saab, N., & Admiraal, W. (2021). Assessment of cognitive, behavioral, and affective learning outcomes in massive open online courses: A systematic literature review. Computers & Education, 163, 104097.

[5] Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart?. Assessment in Education: Principles, policy & practice, 24(3), 317-350.

[6] Memarian, B., & Doleck, T. (2024). A review of assessment for learning with artificial intelligence. Computers in Human Behavior: Artificial Humans, 2(1), 100040.

[7] Tempelaar, D., Nguyen, Q., & Rienties, B. (2020). Learning analytics and the measurement of learning

engagement. Adoption of data analytics in higher education learning and teaching, 159-176.

[8] Santoso, L. W. (2017). Data warehouse with big data technology for higher education. Procedia Computer Science, 124, 93-99.

[9] Harerimana, G., Jang, B., Kim, J. W., & Park, H. K. (2018). Health big data analytics: A technology survey. Ieee Access, 6, 65661-65678.

[10] Lv, X., & Li, M. (2021). Application and research of the intelligent management system based on internet of things technology in the era of big data. Mobile Information Systems, 2021(1), 6515792.

[11] Brock, V., & Khan, H. U. (2017). Big data analytics: does organizational factor matters impact technology acceptance?. Journal of Big Data, 4, 1-28.

[12] Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: applications, prospects and challenges. Mobile big data: A roadmap from models to technologies, 3-20.

[13] Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., ... & Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. Journal of Evidence‐Based Medicine, 13(1), 57-69.

[14] Tang, J. (2021). English teaching ability evaluation algorithm based on big data fuzzy k-means clustering. In Application of Big Data, Blockchain, and Internet of Things for Education Informatization: First EAI International Conference, BigIoT-EDU 2021, Virtual Event, August 1–3, 2021, Proceedings, Part I 1 (pp. 282-291). Springer International Publishing.

[15] Zhang, Y. (2021, September). English teaching evaluation system based on big data. In 2021 4th International Conference on Information Systems and Computer Aided Education (pp. 1321-1325).

[16] Gan, S. (2022, July). Evaluation Model of College English Teaching Effect Based on Big Data Platform. In EAI International Conference, BigIoT-EDU (pp. 512-522). Cham: Springer Nature Switzerland.

[17] Chang, H. (2021). College English flipped classroom teaching model based on big data and deep neural networks. Scientific Programming, 2021(1), 9918433.

[18] Qian, R., Sengan, S., & Juneja, S. (2022). English language teaching based on big data analytics in augmentative and alternative communication system. International Journal of Speech Technology, 25(2), 409-420.

[19] Sun, M., & Li, Y. (2020). Eco-environment construction of English teaching using artificial intelligence under big data environment. IEEE Access, 8, 193955-193965.

[20] Jing, Y., Mingfang, Z., & Yafang, C. (2022). Evaluation model of college English education effect based on big data analysis. Journal of Information & Knowledge Management, 21(03), 2250046.

[21] Tan, Q. (2021, April). Evaluation system of college English teaching based on big data. In Journal of Physics: Conference Series (Vol. 1852, No. 2, p. 022014). IOP Publishing.

[22] Wang, Q., & Jiang, X. Q. (2018, February). Empirical study on reform model of college English teaching model based on computer and big data. In 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 412-415). IEEE.

[23] Wang, C. (2021). Analysis of students' behavior in english online education based on data mining. Mobile Information Systems, 2021(1), 1856690.