Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 9648534, 16 pages
https://doi.org/10.1155/2022/9648534

Research Article

Difficulty-Based Knowledge Point Clustering Algorithm Using Students' Multi-Interactive Behaviors in Online Learning

Zhaoyu Shou,[1] **Jun-Li Lai**,[1] **Hui Wen**,[1] **Jing-Hua Liu**,[1] **and Huibing Zhang**[2]

[1]*School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China*
[2]*School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China*

Correspondence should be addressed to Hui Wen; huiwen@guet.edu.cn

Received 18 April 2022; Revised 12 September 2022; Accepted 13 September 2022; Published 4 November 2022

Academic Editor: Xingling Shao

To improve learners' performance in online learning, a teacher needs to understand the difficulty of knowledge points learners of different cognitive encounter levels in the learning process. This paper proposes a difficulty-based knowledge point clustering algorithm based on collaborative analysis of multi-interactive behaviors. Firstly, combining the group-directed learning path network, forgetting factors and the degree of student-system interaction, we propose a measurement model to calculate the similarity of the difficulty between knowledge points on student-system interactive behavior. Secondly, to solve the data sparsity problem of interaction, we propose an improved similarity model to calculate the similarity of the difficulty between knowledge points on student-teacher and student-student interactive behavior. Finally, the knowledge point difficulty similarity matrix is obtained by integrating the difficulty similarity of knowledge points obtained from student-system interactive behavior, student-teacher interactive behavior, and student-student interactive behavior. The spectral clustering algorithm is used to achieve knowledge point difficulty classification based on the obtained similarity matrix. The experiments on real datasets show that the proposed method has better knowledge point difficulty classification results than the existing methods.

1. Introduction

The development of Internet technology has driven the networking and internationalization of education. Online learning has gained unprecedented attention during COVID-19 prevention and control, which implies it will confront greater chances and problems [1]. Unlike in traditional classrooms, teachers cannot observe the status of students studying knowledge point videos in online learning, making it impossible for teachers to accurately determine the knowledge difficulties that online learners encounter in the learning process. Although teachers can estimate the difficulty of knowledge points empirically, several studies have shown that it is difficult for teachers to determine the correct difficulty level of knowledge points based on the cognitive level of students [2]. There are three main ways to understand the difficulty of knowledge points in the traditional teaching methods: questionnaire surveys, questions, and face-to-face talking. However, these traditional methods are time-consuming and influenced by learners' subjective emotions, resulting in a failure to accurately reflect the real difficulty of the knowledge points.

With the widespread use of online learning platforms such as Coursera, edX, and Udacity, these online learning platforms store the learners' interactive behaviors data in the process of online learning [3, 4]. The literature [5] summarizes online interactive behaviors into three types: student-system interaction (the behavior of learners interacting with knowledge point videos, courses, and resource systems), student-teacher interaction, and student-student interaction. These interactive behaviors reflect the students' learning situations and understanding of knowledge points [6, 7]. For example, when students study a knowledge point more frequently, it means that the knowledge point is more difficult to understand. In terms of knowledge point difficulty, we can automatically obtain the knowledge point clusters at different levels of difficulty through Educational Data Mining (EDM) based on student learning behaviors, which can assist teachers in continuously optimizing the design of teaching content to provide a reliable guarantee for improving teaching efficiency. However, due to factors such as the difficulty of the knowledge points and learning habits, the interactive behavior (student-system, student-teacher, student-student) of different students in learning the knowledge points is different. The challenge of this research paper is how to use the multi-interactive behavior data (student-system, student-teacher, and student-student) in the online education system to accurately classify the difficulty of knowledge points for learners. From the perspective of student-system interactive behavior, the increased frequency and duration of watching videos means that students study difficult knowledge points [8, 9]. From the perspective of learning paths, students' repeated study of the difficult knowledge points forms the structure of the learning sequence of the partial knowledge point cycle [10, 11]. Furthermore, the learning path must take into account students' forgetting behavior [12], which is because students may study knowledge point videos because they forget the content. From the perspective of student-teacher interaction and student-student interactive behaviors, students interact more with their teachers and peers when they study difficult knowledge points. Therefore, the relationship between the different categories of interactive behaviors and the difficulty of the knowledge point needs to be analyzed.

Based on the above analysis, this paper innovatively proposes a difficulty-based knowledge point clustering algorithm using students' multi-interactive behaviors (MIBKPC), which mainly includes the following contributions:

(1) We propose a difficulty-based knowledge point classification algorithm that combines three interactive behaviors to measure the similarity of knowledge point difficulty and provides a more accurate classification of knowledge point difficulty.

(2) Based on student-system interaction, a similarity measurement approach for the difficulty of knowledge points is proposed. The approach integrates the group learning path network, the degree of student-system interaction, and the forgetting behavior of learners, which can assist in measuring the learning process more accurately.

(3) Due to the sparsity problem of student-teacher and student-student interactions, traditional methods of measuring similarity are inaccurate. We propose a similarity measurement method of knowledge point difficulty based on interaction, which resolves the problem by considering all information about student interactions with knowledge points.

The rest of the paper is organized as follows: Section 2 introduces the research related to difficulty-based knowledge point classification. Section 3 describes the definition and calculations related to the algorithm. In Section 4, a difficulty-based knowledge point clustering algorithm using students' multi-interactive behaviors is proposed. Section 5 presents a comparative analysis of different experimental results to evaluate the performance of the proposed algorithm. Finally, Section 6 concludes the work.

2. Related Works

Online learning platforms and educational institutions store a variety of student data. According to the literature [13], Multidimensional data analysis of learning behaviors using educational data mining techniques can help teachers and researchers better understand the learners' learning process. However, only a few studies have been conducted to cluster or classify the difficulty of knowledge points based on the interactive behavior of learners.

In the study of the classification of teaching resources based on statistical methods: Li et al. [8] found that rewatching the video, speeding down, frequent pauses, and skips implied that the videos were more difficult for learners to study Sluis et al. [14] explored the relationship between video complexity and dwelling time or dwelling rate by analyzing learners' clickstream tracking data Brinton et al. [10] proposed an event sequence-based framework to extract repeated subsequences of student behavior to identify recurrent viewing behaviors and found that subsequences were significantly correlated with learning effects Zhu et al. [15] investigated the weight coefficients of the impact of implicit video feedback in the student-system interactive behavior on the learning effect. The implicit video feedback included the video learning frequency, the video learning duration, and the video pausing and dragging frequency. The above works only explore the relationship between the difficulty of videos and interactive behaviors; however, they do not propose a specific method for classifying the difficulty of knowledge point videos. But these works provide a reference for extracting features of interactive behavior as input for knowledge point difficulty classification.

In the study of the classification of teaching resources based on machine learning: Kastrati et al. [16] proposed a video classification framework based on video content by converting the video into text, converting the text into vector space using representation techniques, and training the video classifier using the extracted vectors Othman et al. [17] proposed a classification framework based on video metadata, whereby XML technology was used to extract metadata related to videos, such as video description information and comments information, and then the metadata was used to classify videos using data mining techniques. In the above works, the classification is based on the content of the videos. However, video information only describes the basic content of the video and cannot be used to measure the difficulty of the video. Therefore, these methods cannot be used to classify the difficulty of knowledge point videos.

Zhang et al. [11] proposed a personalized classification algorithm for MOOC videos, which clusters students by their knowledge level, mines the VLBP structure of each class based on their video viewing data by using process mining techniques, and then measures the difficulty and importance of MOOC videos Zhang et al. [9] proposed a difficulty-based clustering method for SPOC videos, using the SimRank++ algorithm to calculate the difficulty similarity between two videos, and then a spectral clustering algorithm is used to achieve video clustering. Though the above algorithms studied the mapping model between the difficulty of videos and the learning behavior, they did not consider the mechanism of the intrinsic association between the multiple interactive behaviors of the learners and the difficulty of knowledge point videos.

To sum up, there is no existing research on knowledge point video difficulty classification based on learners' multidimensional interactive behaviors. In this paper, student-system interactive behaviors, student-teacher interactive behaviors and student-student interactive behaviors in the online learning process are modeled to obtain the knowledge point difficulty similarity matrix, which is combined with the group-directed learning path network and knowledge point difficulty similarity measurement, and then the spectral clustering algorithm is used to classify the knowledge point difficulty.

3. Correlation Definition

In this section, the proposed algorithm's relevant definitions and computational methods are described, and some of the definitions are analyzed and illustrated.

3.1. Knowledge Point. Knowledge point videos refer to the orderly short instructional videos recorded by teachers, which are numbered by the researcher, according to the course knowledge framework. Since the videos are short instructional videos, most of the videos only contain one knowledge point. Therefore, in this paper, one video represents one knowledge point.

3.2. The Degree of Student-System Interaction. The degree of student-system interaction refers to the degree of learners watching knowledge point videos [15]. The recordings of learners watching the knowledge point videos are stored in the online learning platform. From these records, we extracted three behavioral features: the knowledge point video learning frequency, the knowledge point video learning duration, and the knowledge point video pausing and dragging frequency. The calculation process is as follows:

$$SC_{u,i} = \lambda_1 \times f_{sc_{u,i}} + \lambda_2 \times t_{sc_{u,i}} + \lambda_3 \times p_{sc_{u,i}}, \tag{1}$$

where $f_{sc_{u,i}}$ indicates the frequency of student u studying knowledge point i in student-system interactive behavior, $t_{sc_{u,i}}$ represents the duration of student u studying knowledge point i in student-system interactive behavior, and $p_{sc_{u,i}}$ indicates the frequency of pausing and dragging of student u studying knowledge point i in student-system interactive behavior. To unify the units and scales of variables, $f_{sc_{u,i}}$, $t_{sc_{u,i}}$, and $p_{sc_{u,i}}$ are normalized to $[0, 1]$. λ_1, λ_2, and λ_3 are weighting factors. According to literature [15], the best portrayal of the degree of student-system interaction is obtained when $(\lambda_1, \lambda_2, \lambda_3) = (1, 5, 4)$. Then, the student-system interaction degree matrix $SC = [SC_{u,i}]_{m\times n}$ of students can be obtained.

3.3. The Degree of Student-Teacher Interaction. The degree of student-teacher interaction refers to students communicating with the teacher in the online learning process [18]. The online learning platforms store text information that teachers answer questions about knowledge points for students. Extracting keywords from the text and matching them with the knowledge point name allows us to determine which knowledge point is being asked about, then count the effective time and frequency of communication between teachers and students. The degree of student-teacher interaction is portrayed by the frequency and duration of interaction in student-teacher interactive behaviors. The calculation process is as follows:

$$ST_{u,i} = \eta_{st_{u,i}} \times f_{st_{u,i}}, \tag{2}$$

$$\eta_{st_{u,i}} = \begin{cases} \dfrac{t_{st_{u,i}}}{\max\left\{t_{st_{1,i}}, t_{st_{2,i}}, \cdots, t_{st_{m,i}}\right\}}, & \max\left\{t_{st_{1,i}}, t_{st_{2,i}}, \cdots, t_{st_{m,i}}\right\} \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $t_{st_{\mathrm{u,i}}}$ indicates the duration of student-teacher interaction of student u to knowledge point i, $\eta_{st_{u,i}}$ denotes the normalized $t_{st_{\mathrm{u,i}}}$, $\eta_{st_{u,i}}$ takes two values because we consider a possible situation where students have no communication with the teacher during the whole process of a course, which would result in the denominator of 0 for $\eta_{st_{u,i}}$, and $f_{st_{u,i}}$ indicates the frequency of student-teacher interaction of student u to knowledge point i. If student u does not watch for knowledge point i, for such a missing record, we set $SC_{u,i} = 0$. If student u does not ask the teacher a question about knowledge point i in online learning platforms, for such a missing record, we set $ST_{u,i} = 0$. Then, the student-teacher interaction degree matrix $ST = [ST_{u,i}]_{m\times n}$ of students can be obtained.

3.4. The Degree of Student-Student Interaction. The degree of student-student interaction refers to the communication and discussion between students about knowledge points [18]. The online learning platforms also record text records from student-to-student discussions about knowledge points. We can also determine which knowledge point is discussed, such as teacher-student interaction, then count the effective time and frequency of communication between students. It is portrayed by the frequency and duration of student-student interactive behaviors. The degree of student-student interaction is calculated from the following equations:

$$SS_{u,i} = \frac{\sum_{v=1}^{m-1} \eta_{ss_{uv,i}} \times f_{ss_{uv,i}}}{m-1}, \quad u \neq v, \tag{4}$$

$$\eta_{ss_{uv,i}} = \begin{cases} \dfrac{t_{ss_{uv,i}}}{\max\left\{t_{ss_{u1,i}}, t_{ss_{u2,i}}, \cdots, t_{ss_{um,i}}\right\}}, & \max\left\{t_{ss_{u1,i}}, t_{ss_{u2,i}}, \cdots, t_{ss_{um,i}}\right\} \neq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $f_{ss_{uv,i}}$ represents the frequency of interaction between student u and student v to knowledge point i, m denotes the number of students, $t_{ss_{uv,i}}$ indicates the duration of student-student interaction between student u and student v to knowledge point i, $\eta_{ss_{uv,i}}$ denotes the normalized $t_{ss_{uv,i}}$, and $\eta_{ss_{uv,i}}$ takes two values because we consider a possible situation where students communicate with other learners through other tools during the whole process of a course and this communication is not recorded in the learning platform, which would result in a denominator of 0 for $\eta_{ss_{uv,i}}$. If student u does not discuss the knowledge point i with other learners in online learning platforms, for such a missing record, we set $SS_{u,i} = 0$. Then, the student-student interaction degree matrix $SS = [SS_{u,i}]_{m\times n}$ of students can be obtained.

3.5. Directed Learning Path Network. A directed learning path network (DLPN) is a topological network generated based on time series data of student-system interactive behavior [19, 20]. Directed learning path networks can be divided into personal directed learning path networks (PDLPN) and group-directed learning path networks (GDLPN). $\text{PDLPN}_u = (V_u, E_u, W_u)$ represents personal directed learning path networks of student u. $V_u = \{v_1, v_2, \ldots, v_n\}$ indicates the set of knowledge nodes that student u studies, n indicates the number of knowledge points that student u studies. E_u indicates the set of directed edges, and the direction between knowledge nodes indicates the temporal order in which student studies knowledge points, i.e., if student u studies knowledge v_i and then studies knowledge point v_j, the directed edges point from v_i to v_j. W_u is the weight matrix of the learning path, which can be shown in the following formula:

$$W_u = \left[w_{ij}\right]_{n\times n}, \tag{6}$$

where w_{ij} denotes the number of times student u studies from v_i to v_j. GDLPN has the same structure as PDLPN and is defined as $\text{GDLPN}_g = (V_g, E_g, W_g)$. GDLPN is obtained by PDLPN superposition, namely, $W_g = \sum W_u$.

4. Difficulty-Based Knowledge Point Clustering Algorithm Using Multi-Interactive Behaviors (MIBKPC)

The flow block diagram of the difficulty-based knowledge point clustering algorithm using students' multi interactive behaviors (MIBKPC) is shown in Figure 1.

First, the student-system interaction data are analyzed to obtain the SC matrix and GDLPN, and the GDLPN is analyzed to obtain the In-Degree centrality of knowledge node (defined by formula (9)). In-Degree centrality of knowledge node and SC are used as the input of the knowledge point difficulty similarity model based on student-system interaction to obtain the SC-based knowledge point difficulty similarity matrix.

Second, ST matrix and SS matrix are obtained by analyzing the student-teacher interaction and student-student interaction data. ST and SS are used as the input of the knowledge point difficulty similarity model based on interaction to obtain the ST-based knowledge point difficulty similarity matrix and SS-based knowledge point difficulty similarity matrix, respectively.

Finally, the knowledge point difficulty similarity matrix is obtained by linear combination analysis of SC-based knowledge point difficulty similarity matrix, ST-based knowledge point difficulty similarity matrix, and SS-based knowledge point difficulty similarity matrix. A spectral clustering algorithm is used to implement difficulty-based knowledge points clustering based on the obtained similarity matrix.

The proposed algorithm is composed of four parts: the knowledge point difficulty similarity model based on student-system interaction, the knowledge point difficulty similarity model based on interaction, measurement of the difficulty similarity of knowledge points, and spectral clustering based on the difficulty of knowledge points. We have a detailed introduction in the following sections.

4.1. Knowledge Point Difficulty Similarity Model Based on Student-System Interaction. By analyzing the student-system interactive behavior, the knowledge point difficulty similarity is measured only by the degree of student-system interaction that can cause dimension curse. The dimension of the interaction vector of knowledge points in the student-system interaction matrix *SC* increases with the number of learners. When the interaction vectors of knowledge points are used to calculate the difficulty similarity of knowledge points, those pairwise similarities are calculated in high dimensions, which can lead to the problem that the difficulty similarity among knowledge points tends to be the same. The problem can be solved by combining similarity measurements based on the degree of student-system interaction and similarity measurements based on the structure [21, 22]. Therefore, we innovatively construct a knowledge point difficulty similarity model combining the degree of student-system interaction and GDLPN. The model is shown in the following formula:

$$\mathrm{Sim}_{\mathrm{SC}}^{\mathrm{Proposed}}(i,j)=\mathrm{Sim}_{\mathrm{degree}}(i,j)\times\mathrm{Sim}_{\mathrm{GDLPN}}(i,j), \tag{7}$$

where $\mathrm{Sim}_{\mathrm{degree}}(i,j)$ denotes difficulty similarity based on the degree of student-system interaction between knowledge point i and knowledge point j and $\mathrm{Sim}_{\mathrm{GDLPN}}(i,j)$ denotes difficulty similarity based on GDLPN between knowledge point i and knowledge point j.

4.1.1. Knowledge Point Difficulty Similarity Based on the Degree of Student-System Interaction. In online learning, learners usually study each knowledge point video, making the SC matrix a dense matrix. Compared to other similarity approaches in dense matrices, the Adjusted Cosine similarity method can better measure the knowledge point difficulty similarity [23]. This is because learners have different learning preferences (different learning habits and learning foundations), which lead to different interactive behaviors for each learner at the same difficulty level of knowledge point. The Adjusted Cosine similarity method removes the effect of learner preference on the difficulty similarity between two knowledge points by measuring the angle between two decentered knowledge point vectors. Therefore, the knowledge point difficulty similarity measurement based on the degree of system interaction we choose is Adjusted Cosine, as given in the following formula:

$$\mathrm{Sim}_{\mathrm{degree}}(i,j)=\frac{\sum_{u\in KP_i^{\mathrm{SC}}\cap KP_j^{\mathrm{SC}}}\left(SC_{u,i}-\overline{SC_u}\right)\times\left(SC_{u,j}-\overline{SC_u}\right)}{\sqrt{\sum_{u\in KP_i^{\mathrm{SC}}\cap KP_j^{\mathrm{SC}}}\left(SC_{u,i}-\overline{SC_u}\right)^2}\sqrt{\sum_{u\in KP_i^{\mathrm{SC}}\cap KP_j^{\mathrm{SC}}}\left(SC_{u,j}-\overline{SC_u}\right)^2}}, \tag{8}$$

where $SC_{u,i}$ represents the degree of student-system interaction (defined by (1)), $\overline{SC_u}$ represents the average degree of student-system interaction of student u, KP_i^{SC} and KP_j^{SC} represent the set of students who have interacted with knowledge point i and knowledge point j in SC, respectively, and $KP_i^{SC}\cap KP_j^{SC}$ denotes the set of learners interacting with knowledge point i and knowledge point j in common.

4.1.2. Knowledge Point Difficulty Similarity Based on GDLPN. The GDLPN's partial directed learning path diagram is shown in Figure 2. Nodes represent knowledge point videos. The direction of the edge represents the order in which learners study the knowledge point videos. The weight of the edge indicates the number of learning times. The In-Degree of nodes indicates the process of students repeatedly learning knowledge points in GDLPN. In addition, the forgetting curve proposed by Murre and Dros [12] suggests that the students' interactive behaviors that are too far apart in the course learning sequence may be due to the student forgetting the content. To quantify this process, referencing the In-Degree centrality in directed weighted networks [24, 25], this paper integrates the weights of the In-Degree edges, the number of connected In-Degree nodes, and the forgetting distance of students to obtain In-Degree centrality of knowledge node, which is shown in the following formula:

$$\begin{cases} D_{in\_\beta}(v_i)=k_{v_i}^{in\_\beta}\times\left(\dfrac{s_{v_i}^{in\_\beta}}{k_{v_i}^{in\_\beta}}\right)^{\alpha}, \\ s_{v_i}^{in\_\beta}=\displaystyle\sum_{j\in\Gamma_i^{\beta}}\frac{w_{ij}}{d_{ij}}. \end{cases} \tag{9}$$

In formula (9), α is an adjustable parameter, when $0<\alpha<1$, having a large number of connected nodes, is perceived as favorable, whereas when $\alpha>1$, having a small number of connected nodes, is perceived as favorable. β is an adjustable parameter and indicates the forgetting distance of the course learning sequence between two knowledge points. $k_{v_i}^{in\_\beta}$ denotes the number of nodes in the set of directly connected In-Degree nodes within β distance with knowledge node v_i. Γ_i^{β} is the set of directly connected In-Degree nodes within β distance with knowledge node v_i, d_{ij} is the distance of the course learning sequence between knowledge node v_i and knowledge node v_j. w_{ij} is the weight of the edge. We have obtained the highest correct clustering accuracy when $(\alpha,\beta)=(0.6,4)$ after various combinations of α and integers of β.

For example, in Figure 2, the set of directly connected In-Degree nodes with knowledge point 1 is $\{2,3,5,6\}$. Assume $(\alpha,\beta)=(0.6,4)$, the set of directly connected In-Degree

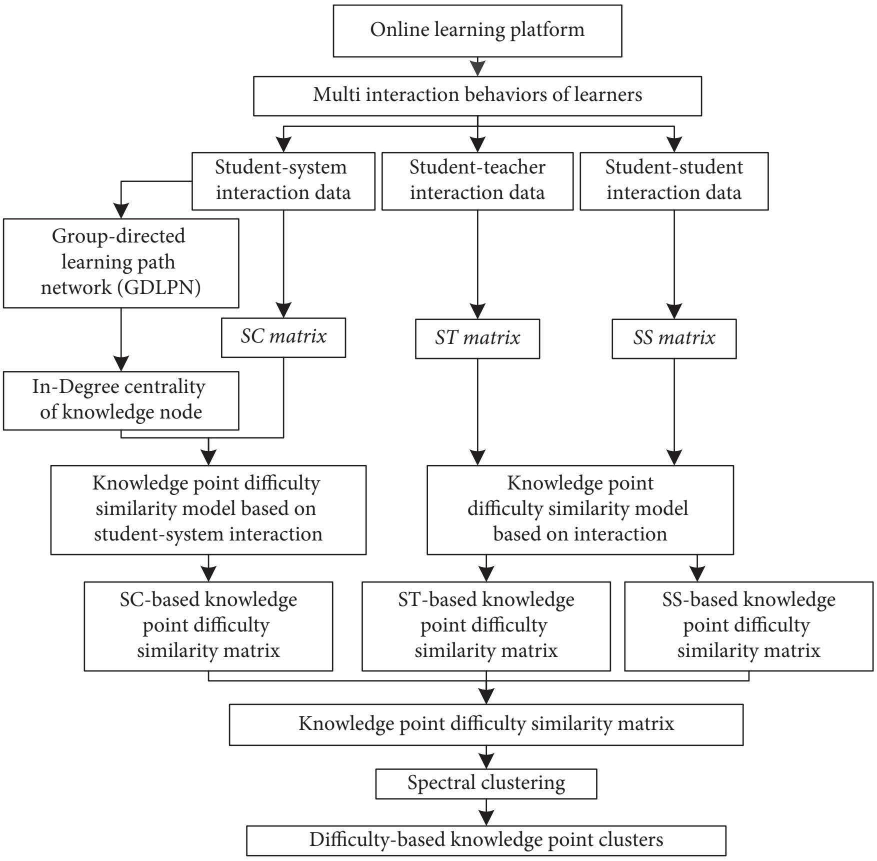

Online learning platform
Multi interaction behaviors of learners
Student-system interaction data
Student-teacher interaction data
Student-student interaction data
Group-directed learning path network (GDLPN)
SC matrix
ST matrix
SS matrix
In-Degree centrality of knowledge node
Knowledge point difficulty similarity model based on student-system interaction
Knowledge point difficulty similarity model based on interaction
SC-based knowledge point difficulty similarity matrix
ST-based knowledge point difficulty similarity matrix
SS-based knowledge point difficulty similarity matrix
Knowledge point difficulty similarity matrix
Spectral clustering
Difficulty-based knowledge point clusters

Figure 1: Block diagram of MIBKPC.

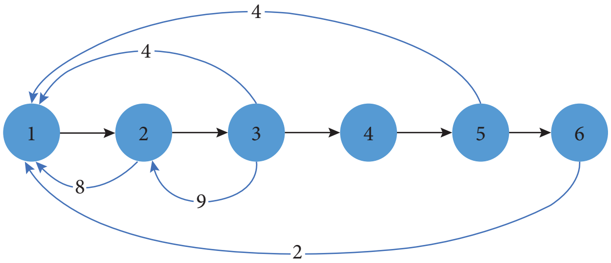


Figure 2: Partial diagram of directed learning path in GDLPN.

nodes within 4 distances with knowledge point 1 is {2, 3, 5}, namely, $\Gamma_1^4 = \{2, 3, 5\}$. $k_{v_1}^{in-4}$ is the number of the set {2, 3, 5}, namely, $k_{v_1}^{in-4} = 3$; similarly, $s_{v_1}^{in-4}$ is the sum of the edge weights of the set {2, 3, 5}, i.e., $s_{v_1}^{in-4} = 8/(2-1) + 4/(3-1) + 4/(5-1) = 11$, and the In-Degree centrality of knowledge point 1 is $D_{in\_4}(v_1) = 3 \times (11/3)^{0.6} \approx 6.54$.

Based on the above analyses, the knowledge point difficulty similarity based on GDLPN is shown in the following formula:

$$\mathrm{Sim}_{\mathrm{GDLPN}}(i, j) = \frac{1}{1 + |D_{in\_\beta}(v_i) - D_{in\_\beta}(v_j)|}, \tag{10}$$

where $D_{in\_\beta}(v_i)$ and $D_{in\_\beta}(v_j)$ denote the In-Degree centrality of knowledge point i and knowledge point j, respectively. From the equation, we can see that the smaller the difference of In-Degree centrality of the two knowledge points, the higher the difficulty similarity of the two knowledge points.

4.2. Knowledge Point Difficulty Similarity Model Based on Interaction.

The ST and SS matrix exist a sparsity problem due to the interaction environment and the teaching model. Traditional difficulty similarity calculation cannot be performed when two knowledge points do not have a co-learner [26, 27]. To solve this problem, we innovatively propose an improved JMSD [28] similarity model through the knowledge point popularity difference and the average interaction degree difference of knowledge points. The specific process is as follows:

Firstly, the popularity of knowledge point i is shown in the following formula:

$$\mathrm{KPD}_i = \frac{|KP_i|}{m}, \tag{11}$$

where $|KP_i|$ indicates the number of learners who have interacted with the knowledge point i, m denotes the number of learners.

Secondly, by analyzing teacher-student interaction and student-student interactive behaviors, we find that the smaller the value of $|\mathrm{KPD}_i - \mathrm{KPD}_j|$, the higher the difficulty similarity of the two knowledge points. $\overline{ID_i}$ denotes the average interaction degree of knowledge point i. The smaller the value of $|\overline{ID_i} - \overline{ID_j}|$, the higher the difficulty similarity of the two knowledge points. Based on the above analysis, this paper constructs a knowledge point difficulty similarity model based on interaction that can be used to calculate knowledge point difficulty similarity for both student-teacher interaction and student-student interaction. The calculation process is as follows:

$$\mathrm{Sim}_{ST}^{\mathrm{Proposed}}(i,j) = \mathrm{Sim}_{SS}^{\mathrm{Proposed}}(i,j) = \begin{cases} S_1(i,j) \times S_2(i,j), & |KP_i \cup KP_j| \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

$$S_1(i,j) = \frac{1}{1 + |\mathrm{KPD}_i - \mathrm{KPD}_j|} \times \frac{1}{2}\left(1 + \frac{|KP_i \cap KP_j|}{|KP_i \cup KP_j|}\right), \tag{13}$$

$$S_2(i,j) = 1 - |\frac{\overline{ID_i} - \overline{ID_j}| + \sum_{u \in KP_i \cap KP_j}\left(ID_{u,i} - ID_{u,j}\right)^2}{|KP_i \cup KP_j|}, \tag{14}$$

where, since the model applies to both ST and SS, it is next shown when the interaction is ST. $ID_{u,i}$ denotes the degree of student-teacher interaction of learner u to knowledge point i, namely, $ID_{u,i} = ST_{u,i}$, $\overline{ID_i} = \overline{ST_i}$. KP_i and KP_j denote the set of learners who have interacted with knowledge point i and knowledge point j in ST matrix, respectively. KPD_i and KPD_j represent the popularity of knowledge point i and knowledge point j, respectively. $|KP_i \cap KP_j|$ denotes the number of learners interacting with knowledge point i and knowledge point j in common. $KP_i \cup KP_j$ denotes the union of KP_i and KP_j. SS is similar to ST in this way. The traditional JMSD uses only the co-learner interaction data to portray the difficulty similarity, the proposed similarity model makes full use of the interaction data of the two knowledge points.

4.3. Measurement of the Difficulty Similarity of Knowledge Points. Considering this situation, communication between students and teachers and between students and students may not occur throughout the course. To solve this problem, the difficulty similarity between two knowledge points is obtained by weighting the combination of $\mathrm{Sim}_{SC}^{\mathrm{Proposed}}(i,j)$, $\mathrm{Sim}_{ST}^{\mathrm{Proposed}}(i,j)$, and $\mathrm{Sim}_{SS}^{\mathrm{Proposed}}(i,j)$ (defined by equation (12)). $\mathrm{Sim}_{\mathrm{DKP}}^{\mathrm{Proposed}}(i,j)$ can be calculated even when both $\mathrm{Sim}_{ST}^{\mathrm{Proposed}}(i,j) = 0$ and $\mathrm{Sim}_{SS}^{\mathrm{Proposed}}(i,j) = 0$. The calculation process is shown in the following formula:

$$\begin{cases} \mathrm{Sim}_{\mathrm{DKP}}^{\mathrm{Proposed}}(i,j) = \alpha_1 \times \mathrm{Sim}_{SC}^{\mathrm{Proposed}}(i,j) + \alpha_2 \times \mathrm{Sim}_{ST}^{\mathrm{Proposed}}(i,j) + \alpha_3 \times \mathrm{Sim}_{SS}^{\mathrm{Proposed}}(i,j), \\ \alpha_1 + \alpha_2 + \alpha_3 = 1, \end{cases} \tag{15}$$

where α_1, α_2, and α_3 are weighting factors, which can be adjusted according to the cognitive level of different student groups. Due to $\alpha_1 + \alpha_2 + \alpha_3 = 1$, we only need to get two of the parameters to get the other one. A total of 231 different combinations $(\alpha_1, \alpha_2, \alpha_3)$ were tested under that both α_1 ranging from 0 to 1 with increment of 0.05 and α_2 ranging from 0 to 1 with increment of 0.05, and it was concluded that the clustering accuracy was the highest when the value was $(\alpha_1, \alpha_2, \alpha_3) = (0.8, 0.1, 0.1)$.

4.4. Spectral Clustering Based on the Difficulty of Knowledge Points. Teachers determine the number of knowledge point difficulty clusters in the actual teaching. Assuming N knowledge points, which is divided into K categories, the specific process of the spectral clustering algorithm is as follows:

Firstly, this paper constructs the knowledge point difficulty similarity matrix $M_{i,j} = \mathrm{Sim}_{\mathrm{DKP}}^{\mathrm{Proposed}}(i,j)$, and then the Laplacian matrix L of M is calculated as follows:

$$L = D^{-1/2}(D - M)D^{-1/2}, \tag{16}$$

where D is the diagonal matrix; $D_{ii} = \sum_{j=1}^{N} \mathrm{Sim}_{\mathrm{DKP}}^{\mathrm{Proposed}}(i,j)$.

Then, the eigenvector of the Laplacian matrix L is calculated, and the eigenvector corresponding to the first K minimum eigenvalues is extracted, forming F of $N \times K$ dimension. The K-means algorithm is used to cluster the feature subspace F with the number of K.

The MIBKPC is shown in Table 1.

Table 1: Algorithm MIBKPC (pseudo-code).

```
Algorithm 1 MIBKPC
Input:
  $UD$: learner's multidimensional interactive data set $UD = \{d_u | u = 1, \cdots, m\}$; $m$ is the number of learners
  $N$: number of knowledge point videos
  $K$: number of clusters
Output: $K$ knowledge points clusters
for each $d_u \in U\ D$ do
  According to Section 3.5, construct $(\mathrm{PDLPN}_u)$
  for each $i \in N$ do
    Calculate $SC_{u,i}$ using formula (1)
    Calculate $ST_{u,i}$ using formula (2)
    Calculate $SS_{u,i}$ using formula (4)
  end for
end for
for each $d_u \in UD$ do
  $GPDLPN = \sum PDLPN_u$
end for
for each $i \in N$ do
  for each $j \in N$ do
    Calculate $\mathrm{Sim}_{SC}^{\mathrm{Proposed}}(i, j)$ using formula (7)
    Calculate $\mathrm{Sim}_{ST}^{\mathrm{Proposed}}(i, j)$ using formula (12)
    Calculate $\mathrm{Sim}_{SS}^{\mathrm{Proposed}}(i, j)$ using formula (12)
    Calculate $\mathrm{Sim}_{DKP}(i, j)$ using formula (15)
  end for
Spectral clustering of $K$ classes according to Section 4.4
Return $K$ knowledge points clusters
```

5. Experimental Results and Analysis

In this section, we evaluate the effectiveness of the proposed algorithm (MIBKPC). Firstly, MIBKPC is compared with commonly used classical methods for knowledge point difficulty classification accuracy. Secondly, we have examined the generalizability of the MIBKPC algorithm, which also obtains good classification results by relying only on the student-system interactive behavior. Thirdly, to verify the superiority of MIBKPC on knowledge point difficulty similarity calculation, other similarity methods are compared for clustering precision. Fourthly, the results of knowledge point difficulty classification for learners at different cognitive levels are analyzed. Finally, the relationship between the three interactive behaviors and the difficulty of knowledge points for learners at different cognitive levels is analyzed.

5.1. Data Sets. The data source was obtained from the interactive behavior data of 2019 students participating in the Data Structure and Algorithm course, which is a mandatory course for sophomores at a university. The experimental dataset consists of 207 knowledge point videos, 77,753 video-watching records of 362 students, 8422 text records of interactions between teachers and students, and 1463 text records of interactions between students and knowledge point test data.

5.2. Dataset Preprocessing. In part of the preprocessing of the experiment, learners who watched more than 10 seconds of each video were considered to have learned the knowledge point effectively, so we removed the records of students who watched each video for less than 10 seconds. Furthermore, after deleting the records, the watching records of students who viewed less than 1/3 of all videos are deleted, as well as the corresponding student-teacher interaction and student-student interaction records. After the data preprocessing, we retained 50,544 video-watching records of 272 students, 7683 text records of interactions between teachers and students, and 1252 text records of interactions between students. The duration of knowledge point videos is shown in Figure 3.

5.3. Experimental Evaluation. We use the external evaluation method to evaluate the clustering results [29], and the external evaluation method needs to obtain the real difficulty of knowledge points. We use $\overline{\mathrm{mkp}_i}$ to measure the real difficulty of knowledge points, if $\overline{\mathrm{mkp}_i}$ of a knowledge point is lower than other knowledge points, indicating that the relative real difficulty of the knowledge point is higher. $\overline{\mathrm{mkp}_i}$ is calculated as follows:

Students take a test after completing the course, we define the test score matrix S based on students' scores on each question, namely, $S = \begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{k1} & \cdots & s_{km} \end{bmatrix}$. Teachers analyze the test papers to get the relationship between test questions and knowledge points, and construct the knowledge point test question association matrix $F = \begin{bmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{k1} & \cdots & f_{km} \end{bmatrix}$, which determines whether test

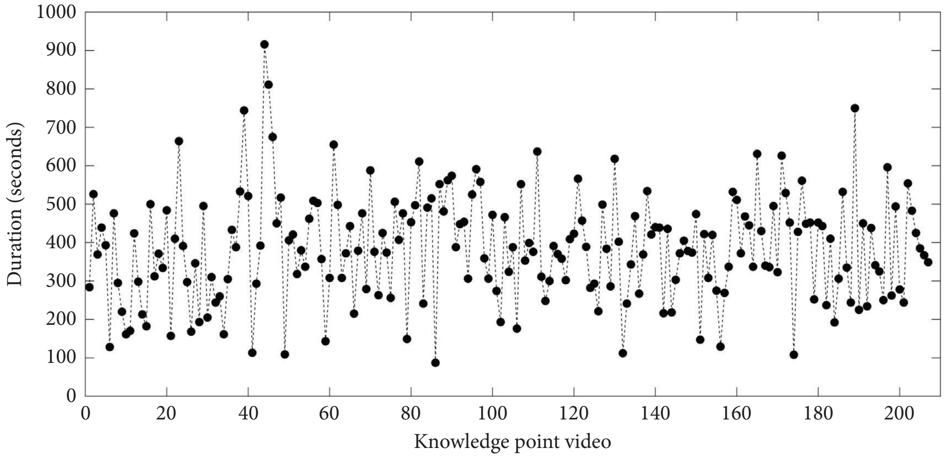


Figure 3: The duration of knowledge point videos.

questions contain knowledge points. The mastery degree of knowledge point is shown in the following formula:

$$\mathrm{mkp}_{u,i} = \frac{\sum_{j=1}^{J} f_{ij} s_{u,j}}{\sum_{j=1}^{J} f_{ij} \mathrm{score}_j}, \tag{17}$$

where J represents the number of test questions to knowledge point i, $s_{u,j}$ represents the test score of student u on test question j, f_{ij} indicates the degree of relevance of knowledge point i to test question j (0-irrelevant, 1-part relevant, 2-indirect relevant, and 3-direct relevant) [18], and score_j represents the score of test question j.

The average mastery degree of knowledge point is shown in the following formula:

$$\overline{\mathrm{mkp}_i} = \frac{\sum_{u=1}^{m} \mathrm{mkp}_{u,i}}{m}, \tag{18}$$

where m is the number of learners.

In addition, a teacher pays more attention to the most difficult and easiest knowledge points for learners to optimize the content design. The effectiveness of the algorithm is measured by the set of the most difficult and easiest knowledge points [9]. The process is as follows:

First, all knowledge points to be ranked in order of $\overline{\mathrm{mkp}_i}$ from lowest to highest, and then the knowledge points are divided into K classes on average to obtain the set of the most difficult knowledge points is defined as D_{mkp} and the set of the easiest knowledge points is defined as E_{mkp}. Taking $K = 3$ as an example, the knowledge points sort from low to high by $\overline{\mathrm{mkp}_i}$, and then the 207 knowledge points are divided into 3 classes on average, then D_{mkp} is the set of the top 69 knowledge points and E_{mkp} is the set of the last 69 knowledge points. Second, based on the knowledge point clusters given by the proposed algorithm, the teacher selects the most difficult and easiest clusters of knowledge points, denoted as D_c and E_c, respectively. For any knowledge point $v \in D_{\mathrm{mkp}}$, md_{kp} denotes the score of the most difficult clustering result. me_{kp} denotes the score of the easiest clustering result. The precision of the clustering result is defined as *PRE*. md_{kp}, me_{kp}, and PRE are shown in the following equation:

$$md_{kp} = \begin{cases} 1, & v \in D_c, \\ -1, & v \in E_c, \\ 0, & v \in V - (D_c \cup E_c), \end{cases} \tag{19}$$

$$me_{kp} = \begin{cases} 1, & v \in E_c, \\ -1, & v \in D_c, \\ 0, & v \in V - (D_c \cup E_c), \end{cases} \tag{20}$$

$$\mathrm{PRE} = \frac{1}{2}\left(\frac{\sum_{v \in D_{\mathrm{mkp}}} md_{kp}}{|D_{\mathrm{mkp}}|} + \frac{\sum_{v \in E_{\mathrm{mkp}}} me_{kp}}{|E_{\mathrm{mkp}}|}\right). \tag{21}$$

According to equation (19), V denotes the set of knowledge points. For a knowledge point in D_{mkp}, if it belongs to D_c, which indicates that the algorithm gives the correct classification, then $md_{kp} = 1$, if it belongs to E_c, which indicates that the algorithm gives the wrong classification, then $md_{kp} = -1$, and if it is not in D_c or E_c, then $md_{kp} = 0$. Similarly, equation (20) gives the score of the knowledge point clustering result in E_{mkp}. Equation (21) is the final precision of the clustering result.

5.4. Experimental Settings.

Based on the selected dataset and evaluation method, we used the proposed algorithm (MIBKPC) to obtain knowledge point clusters and the precision of the algorithm under different numbers of clusters K. We set the number of clusters K as 2, 3, and 5.

5.5. Experimental Results.

In the first experiment, to verify the effectiveness of the proposed algorithm (MIBKPC), we compared it with the three commonly used classical methods. The three commonly used classical methods are defined as follows.

The first method is defined as MS, which sorts the knowledge points in ascending order by interaction degree of knowledge points ikp_i ($ikp_i = \sum_{u=1}^{m} SC_{u,i} + ST_{u,i} + SS_{u,i}$), the knowledge points are divided into K groups on average. The first group with the minimum average interaction degree is the easiest knowledge point set. The last group with the maximum average interaction degree is the most difficult knowledge point set.

The second method is defined as MC, which uses K-means clustering algorithm to cluster the knowledge points based on ikp_i. The cluster of the maximum average interaction degree is the most difficult knowledge point set. The cluster of the minimum average interaction degree is the easiest knowledge point set.

The third method is defined as MVC, which defines a 4-dimensional interaction feature vector $(SC_i, ST_i, SS_i, D_{in\_\beta}(v_i))$. The knowledge points are clustered based on the interaction feature vector by using K-means clustering algorithm. The knowledge point cluster with the maximum average interaction degree is the most difficult knowledge point set, and the knowledge point cluster with the minimum average interaction degree is the easiest knowledge point set.

The knowledge points are classified by MIBKPC, MS, MC, and MVC under the conditions of $K = 2$, 3, and 5, respectively, and the classified results are evaluated to obtain the corresponding precision of the clustering PRE. The experimental results are shown in Figure 4.

Figure 4 shows that for different K values, the clustering precision of the proposed algorithm MIBKPC is higher than that of MS, MC, and MVC. Additionally, the MIBKPC's precision is at its maximum when $K = 3$, which is congruent with the actual teaching experience of teachers, as they often classify the difficulty level of knowledge points into three categories. The analysis of the experimental data showed that some easy knowledge points have a higher average interaction degree than the difficult knowledge points, so MS cannot distinguish the knowledge point difficulty by only relying on the average interaction degree of knowledge points. MC and MVC are easily affected by the individual knowledge points with higher or lower interaction degrees, which leads to inaccurate classification results. Therefore, the MIBKPC algorithm proposes a corresponding model to measure the similarity of the difficulty between two knowledge points according to the behavioral characteristics of different interactions, which can well quantify the learning process of learners and reduce the influence of knowledge points with higher interactions on the clustering results.

In the second experiment, we considered that some learning platforms do not provide functions for student-teacher interaction and student-student interaction, resulting in the inability to collect data on student-teacher interaction and student-student interaction. However, the MIBKPC algorithm requires three types of interactive behavior data to be applied. To verify the generalizability of the MIBKPC algorithm, we define the MIBKPC-SC algorithm which is a simplification of the MIBKPC algorithm. The MIBKPC-SC algorithm requires only the student-system interactive behavior data of learners to achieve knowledge point difficulty clustering. Firstly, MIBKPC-SC measures $\mathrm{Sim}_{\mathrm{SC}}^{\mathrm{Proposed}}(i, j)$ based on student-system interaction data. Then, let $\mathrm{Sim}_{\mathrm{DKP}}^{\mathrm{Proposed}}(i, j) = \mathrm{Sim}_{\mathrm{SC}}^{\mathrm{Proposed}}(i, j)$. The knowledge difficulty clustering is performed according to Section 4.4. The clustering results of MIBKPC and MIBKPC-SC under the conditions of $K = 2$, 3, and 5 are shown in Figure 5.

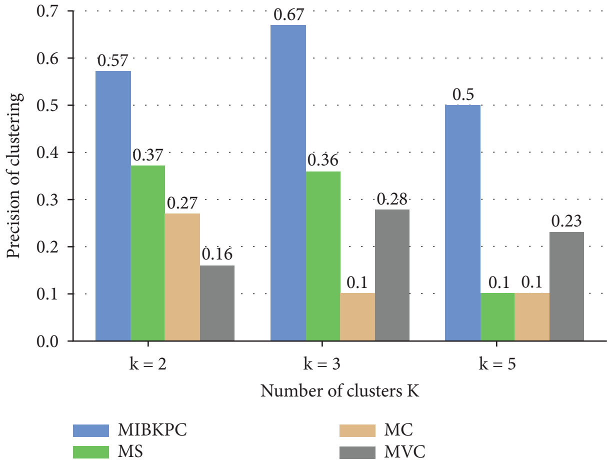


Figure 4: Precision of MIBKPC, MS, MC, and MVC.

From Figure 5, the clustering accuracy of MIBKPC-SC is similar to that of the MIBKPC algorithm when $K = 2$ and $K = 5$. When $K = 3$, the clustering accuracy of MIBKPC-SC decreases by 20 percent compared with that of the MIBKPC algorithm. The analysis of the experimental data revealed that the clustering accuracy of the MIBKPC-SC algorithm is reduced because of the inaccurate classification of easy knowledge points, while MIBKPC can provide a more accurate difficulty similarity portrayal of knowledge points with different difficulty levels by considering three interactive behaviors. Based on the above analysis, MIBKPC-SC has a good effect on the clustering results, although it is not as high as the clustering accuracy of the MIBKPC algorithm. Furthermore, our proposed MIBKPC algorithm framework can be applied to most learning platforms.

In the third experiment, we verify the superiority of MIBKPC in the similarity calculation of knowledge point difficulty. MIBKPC constructs the similarity matrix of knowledge point difficulty from three similarity models $(\mathrm{Sim}_{SC}^{\mathrm{Proposed}}(i, j)$, $\mathrm{Sim}_{ST}^{\mathrm{Proposed}}(i, j)$, and $\mathrm{Sim}_{SS}^{\mathrm{Proposed}}(i, j))$ to perform spectral clustering and thus achieve the difficulty classification of knowledge points. This paper uses other similarity methods for the similarity calculation of knowledge point difficulty. We set $\mathrm{Sim}_{\mathrm{SC}}^{\mathrm{ACOS}}(i, j) = \mathrm{Sim}_{\mathrm{degree}}^{\mathrm{ACOS}}(i, j)$ in the student-system interactive behavior since traditional similarity models do not measure similarity from the perspective of GDLPN. In the case of the Adjusted Cosine similarity (ACOS) [30], the specific process is as follows:

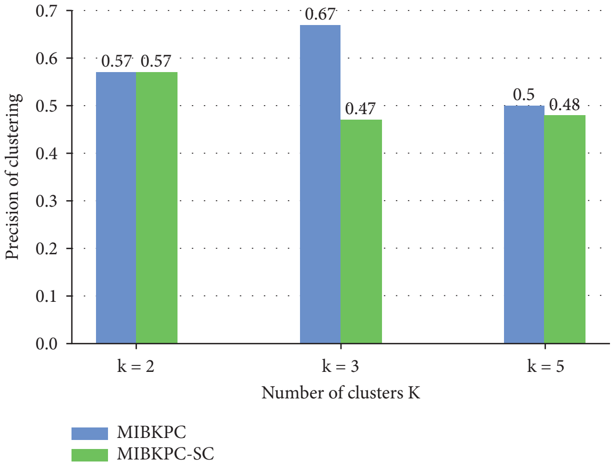


Figure 5: Precision of MIBKPC and MIBKPC-SC.

$$\begin{aligned}\mathrm{Sim}_{\mathrm{DKP}}^{\mathrm{ACOS}}(i,j) &= \alpha_1 \times \mathrm{Sim}_{SC}^{\mathrm{ACOS}}(i,j) + \alpha_2 \times \mathrm{Sim}_{ST}^{\mathrm{ACOS}}(i,j) + \alpha_3 \times \mathrm{Sim}_{SS}^{\mathrm{ACOS}}(i,j) \\ &= \alpha_1 \times \mathrm{Sim}_{\mathrm{degree}}^{\mathrm{ACOS}}(i,j) + \alpha_2 \times \mathrm{Sim}_{ST}^{\mathrm{ACOS}}(i,j) + \alpha_3 \times \mathrm{Sim}_{SS}^{\mathrm{ACOS}}(i,j).\end{aligned} \tag{22}$$

Similarly, this paper compares with SimRank++ [9], RJMSD [26], ACOS [30], PCC [31], and JMSD [28] similarity methods to obtain the corresponding precision of the clustering. The experimental results are shown in Figure 6.

As shown in Figure 6, under the conditions of different K values, the clustering precision of MIBKPC is better than other similarity models. For the traditional similarity models, JMSD and ACOS have better accuracy, and PCC is worse. SimRank++ has better clustering accuracy than RJMSD. Due to their general application to student similarity calculations, PCC and RJMSD are less effective when calculating similarity of knowledge point difficulty. Moreover, SimRank++ and JMSD are based on the structural perspective to portray the difficulty similarity of knowledge points, which can have better clustering accuracy, but there is only co-interaction data that the two algorithms consider. Therefore, the proposed similarity model can more precisely portray the knowledge point difficulty similarity by making full use of the knowledge point interaction data and considering GDLPN and interaction degree.

In the fourth experiment, we analyze the influence of learners at different cognitive levels on the effectiveness of MIBKPC. We divide the dataset into three datasets (primary, intermediate, and advanced) according to the cognitive level of learners, and then obtain the clustering precision of MIBKPC algorithm in Figure 7.

In Figure 7, the clustering precision of MIBKPC decreases on the three datasets compared with the second experiment at different K values. The clustering precision of intermediate learners is the best. To better analyze the reasons for the decrease in clustering precision, this paper further combines GDLPN and knowledge point difficulty clustering results analysis of learners at different cognitive levels; GDLPN is shown in Figure 8. Table 2 shows the final knowledge point difficult clustering result.

In learning path networks, the red line represents the In-Degree edge and the green line represents the Out-Degree edge. From Figures 8(b) and 8(c), we find that intermediate learners and advanced learners repeatedly study the difficult knowledge points, and their interactive behavior data can reflect the difficulty of knowledge points, thus their clustering accuracy is better than that of primary learners. By comparing Figures 8(a) and 8(b), we can see that primary learners watch the knowledge video only once and rarely repeat the knowledge points, which leads to the interactive behaviors not reflecting their real learning effects, resulting in the lowest clustering accuracy. According to Table 2, we can see that advanced learners have the least difficulty with knowledge points, while primary learners have more difficulty than intermediate learners. We find that advanced learners have stronger learning abilities, so even if they encounter difficult knowledge points, they can understand them quickly, which results in fewer difficult knowledge points for advanced learners than for intermediate learners.

In the fifth experiment, based on the three data sets divided in the fourth experiment, we explore the relationship between three interactive behaviors and the difficulty of knowledge points for learners at different cognitive levels. According to formula (15), α_1, α_2, and α_3 represent the importance of student-system interaction, student-teacher interaction, and student-student interaction on the difficulty of the knowledge points, respectively. Due to $\alpha_1 + \alpha_2 + \alpha_3 = 1$, we only need to get two of the parameters to get the other one. Under the conditions of different K values, we obtain the average clustering precision of the MIBKPC algorithm in three groups (primary, intermediate, and advanced) for both α_1 ranging from 0 to 1 with increment of 0.05 and α_2 ranging from 0 to 1 with increment of 0.05. The experimental results are shown in Figure 9.

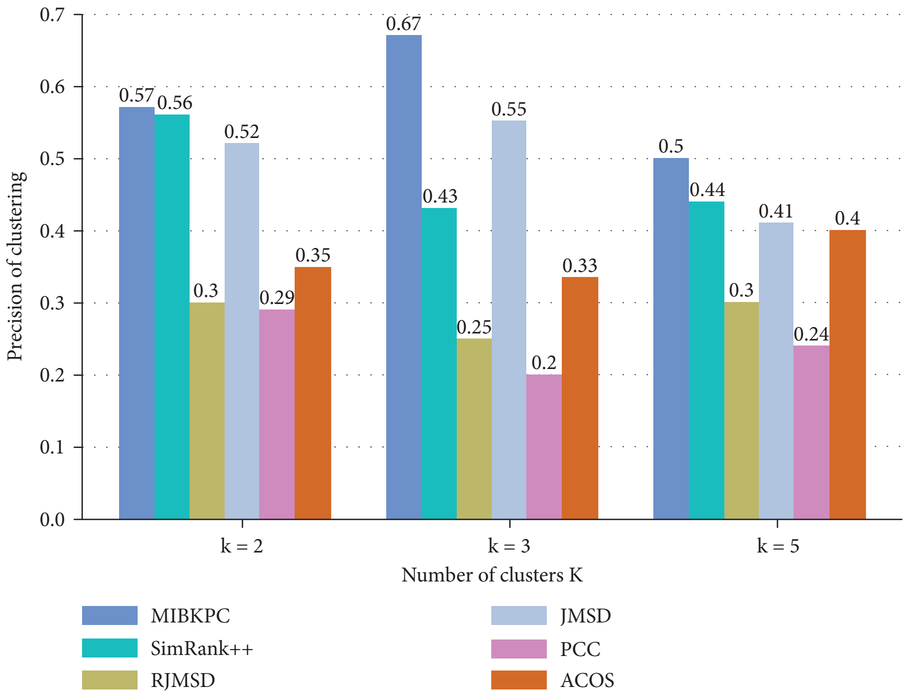


Figure 6: Comparison of clustering precision of MIBKPC with other similarity methods.

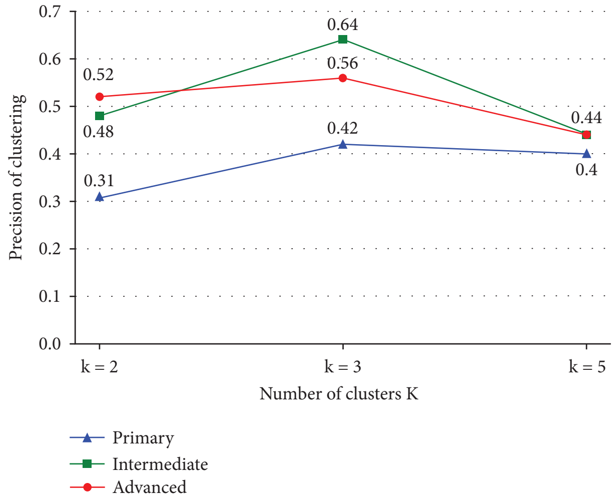


Figure 7: MIBKPC's algorithm precision of students at different cognitive levels.

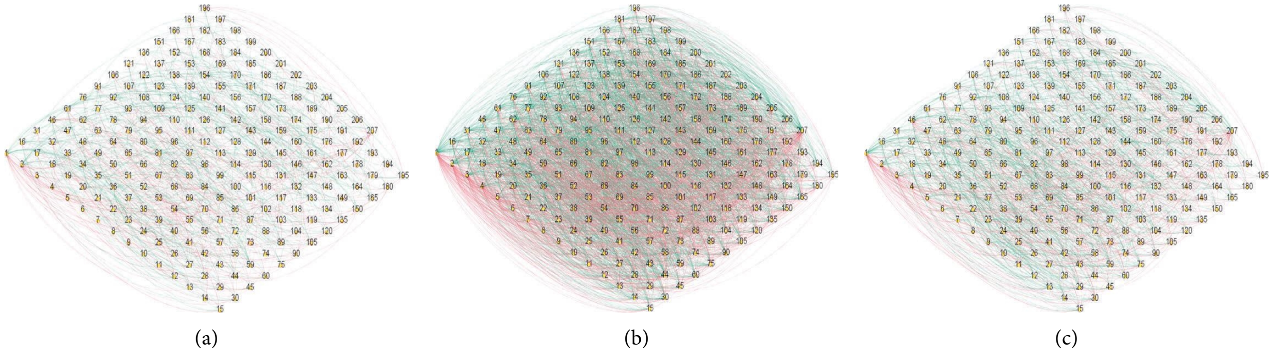
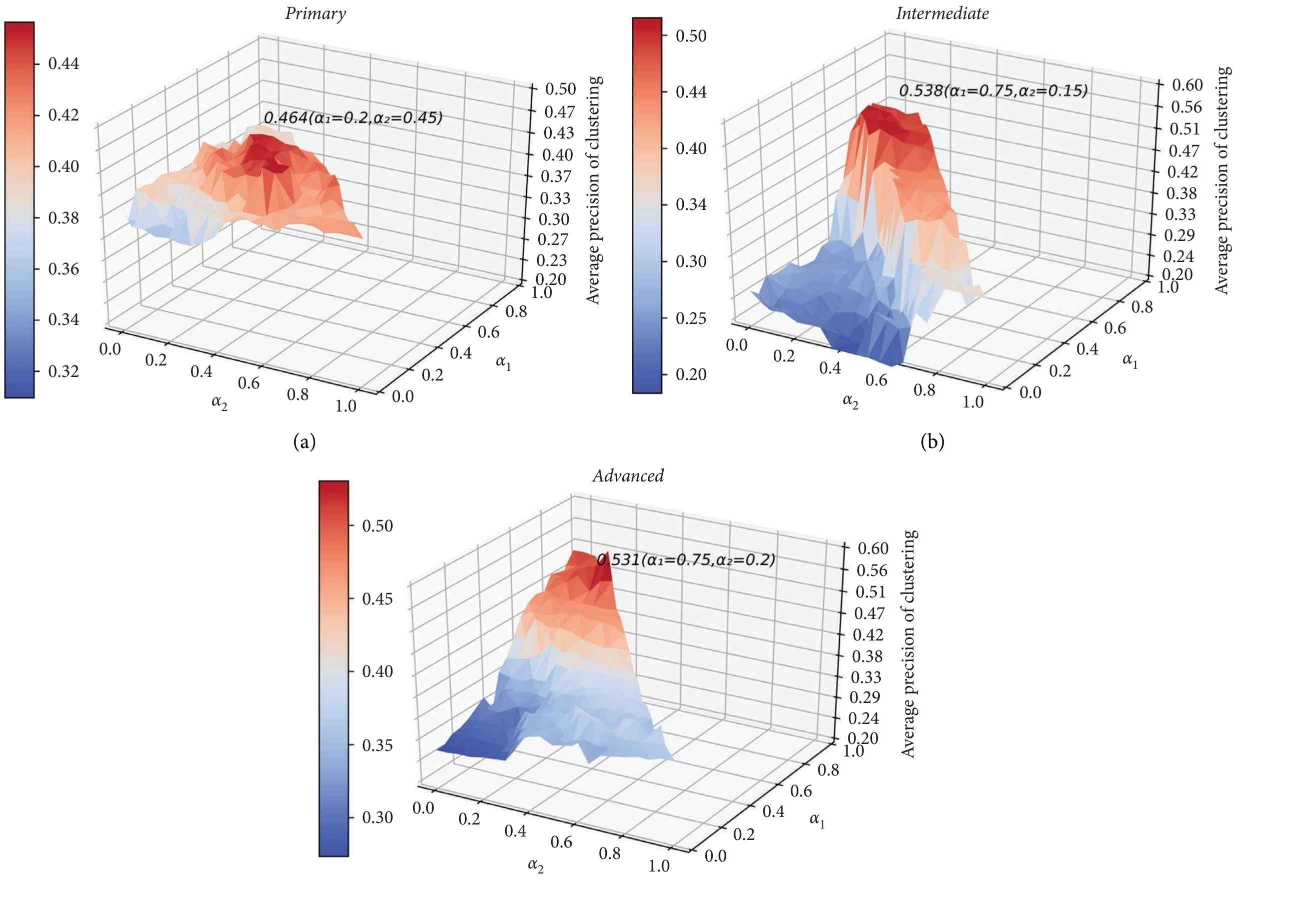

(a)
(b)
(c)

Figure 8: Learning path networks of learners at different cognitive levels. (a) Primary. (b) Intermediate. (c) Advanced.

Table 2: Knowledge point difficulty classification of learners at different cognitive levels.

	Classification by difficult ($K = 3$)		
	Number of learners	Knowledge point name	Number of knowledge points
Primary	69	v_{76}: circular queue in data structure, v_{79}: enqueue and dequeue, and v_{87}: special binary tree, . . .	81
Intermediate	171	v_{119}: deleting priority queue implemented by heap, v_{126}: pre-order traversal of a tree, and v_{127}: post-order traversal of a tree, . . .	74
Advanced	32	v_{126}: pre-order traversal of a tree, v_{127}: post-order traversal of a tree, and v_{129}: storage representation of tree, . . .	71



(a)

(b)

(c)

Figure 9: Clustering precision of learners in three groups for different sizes of α_1 and α_2. (a) Primary. (b) Intermediate. (c) Advanced.

From Figure 9(a), primary learners have the highest clustering accuracy when $(\alpha_1, \alpha_2, \alpha_3) = (0.2, 0.45, 0.35)$, indicating that student-teacher and student-student interactions for primary learners are better indicators of knowledge point difficulty than student-system interactions. $\alpha_1, \alpha_2, \alpha_3$ are different from equation (15), for the dataset of primary learners, we find that primary learners are less engaged while watching the knowledge point videos, which results in the student-system interactive behavior not reflecting the level of difficulty of the knowledge points, thus a smaller weight value of α_1 is assigned. From Figures 9(b) and 9(c), the clustering accuracy of advanced and intermediate learners is gradually increasing with the increase of α_1, which indicates that they invest more learning in the knowledge point videos, and their system interactive behavior can better reflect the knowledge point difficulty. Intermediate learners have the highest clustering accuracy when $(\alpha_1, \alpha_2, \alpha_3) = (0.75, 0.15, 0.1)$. When $(\alpha_1, \alpha_2, \alpha_3) = (0.75, 0.2, 0.05)$, advanced learners have the highest clustering accuracy. It can be inferred that student-teacher interaction of advanced learners can better reflect knowledge point difficulty than that of intermediate learners.

Table 3: Symbols, semantics, and the respective values.

Symbols	Meaning and value
u	Index for the learner that can have values as $1, 2, \ldots m$. m is number of learners
i	Index of knowledge point that can have values as $1, 2, \ldots n$. n is number of knowledge points
$SC_{u,i}$	The degree of student-system interaction associated with learner u and knowledge point i. Values are in the range of 0 to 1
$ST_{u,i}$	The degree of student-teacher interaction associated with learner u and knowledge point i. Values are in the range of 0 to 1
$SS_{u,i}$	The degree of student-student interaction associated with learner u and knowledge point i. Values are in the range of 0 to 1
$f_{sc_{u,i}}$	The frequency of student-system interaction of student u to knowledge point i. Values are in the range of 0 to 1
$t_{sc_{u,i}}$	The duration of student-system interaction of student u to knowledge point i. Values are in the range of 0 to 1
$p_{sc_{u,i}}$	The frequency of pausing and dragging of student u studying knowledge point i. Values are in the range of 0 to 1
λ_1, λ_2, λ_3	Weighting factors associated with $SC_{u,i}$. Here, we considered values as $(\lambda_1, \lambda_2, \lambda_3) = (1, 5, 4)$
PDLPN_u	Directed learning path network associated with learner u
GDLPN_g	Group-directed learning path networks
$D_{in\_\beta}(v_i)$	The in-degree of nodes associated with knowledge point i.
α, β	Adjustable parameters associated with $D_{in\_\beta}(v_i)$. Here, we considered values as $(\alpha, \beta) = (0.6, 4)$
KP_i	The set of learners who have interacted with knowledge point i
KPD_i	The popularity of knowledge point i. Values are in the range of 0 to 1
$ID_{u,i}$	The degree of interaction of learner u to knowledge point i. If in the ST matrix, $ID_{u,i} = ST_{u,i}$. If in the SS matrix, $ID_{u,i} = SS_{u,i}$
$\text{Sim}_{SC}^{\text{Proposed}}(i, j)$	The difficulty similarity based on student-system interaction between knowledge point i and knowledge point j. Values are in the range of 0 to 1
$\text{Sim}_{\text{degree}}(i, j)$	The difficulty similarity based on the degree of student-system interaction between knowledge point i and knowledge point j. Values are in the range of 0 to 1
$\text{Sim}_{\text{GDLPN}}(i, j)$	The difficulty similarity based on group-directed learning path networks between knowledge point i and knowledge point j. Values are in the range of 0 to 1
$\text{Sim}_{ST}^{\text{Proposed}}(i, j)$	The difficulty similarity based on student-teacher interaction between knowledge point i and knowledge point j. Values are in the range of 0 to 1
$\text{Sim}_{SS}^{\text{Proposed}}(i, j)$	The similarity of the difficulty based on student-student interaction between knowledge point i and knowledge point j. Values are in the range of 0 to 1
$\text{Sim}_{\text{DKP}}^{\text{Proposed}}(i, j)$	The difficulty similarity between knowledge point i and knowledge point j. Values are in the range of 0 to 1
$\alpha_1, \alpha_2, \alpha_3$	Weighting factors associated with $\text{Sim}_{\text{DKP}}^{\text{Proposed}}(i, j)$. Here, we considered values as $(\alpha_1, \alpha_2, \alpha_3) = (0.8, 0.1, 0.1)$
K	Number of clusters. Here, we considered values as 2, 3, 5
$\overline{\text{mkp}}_i$	The average mastery degree of knowledge point associated with knowledge point i.
ikp_i	Interaction degree of knowledge point i.
PRE	Precision of the clustering

6. Conclusions and Future Works

This paper proposes a difficulty-based knowledge point clustering algorithm using students' multi-interactive behaviors. Firstly, we propose a knowledge point difficulty similarity model based on student-system interaction. The model innovatively combines interaction degrees and learning paths. Secondly, to solve the problem of the sparsity of student-teacher interaction and student-student interaction, we propose a knowledge point difficulty similarity model based on interactive behavior by using the full information of interaction data. Finally, the knowledge difficulty similarity matrix obtained by three types of interactive behavior is used to obtain the knowledge point difficulty classification using spectral clustering.

The proposed algorithm helps teachers understand the difficult knowledge points of learners for better optimization of teaching process design and teaching content. If an easy knowledge point is always classified into a difficult cluster and the difficulty level of the knowledge point is not the same as the teacher considered, teachers can consider whether the video of the knowledge point is not well explained or enhance the explanation of the knowledge point in class to better optimize the teaching process design and teaching content.

The proposed algorithm can be used with tiny sample datasets, which does not need data to be trained, and can also be applied to some learning platforms that store only student-system interactive behavior data. Based on the analysis of the experimental results, our proposed algorithm has better results in classifying the difficulty of knowledge points compared to other existing methods. If we can collect more student-teacher and student-student interaction data, as sometimes some students interact with each other through other platforms, the algorithm will better measure the difficulty similarity between knowledge points and thus improve the clustering accuracy.

The knowledge point difficulty classification method proposed in this paper is for groups of students. It does not provide individualized knowledge point difficulty classification for each student with different learning preferences. To achieve personalized education with multiple intelligences and improve learners' learning effectiveness, future studies need to examine the relationship between the difficulty of knowledge points and individual learners. Therefore, we should examine whether learners' learning behaviors contain more behavioral characteristics related to the difficulty of knowledge points. We should also develop prediction models for predicting learners' knowledge difficulties, and recommend multiple intelligence learning strategies that meet each student's needs.

Appendix

The symbols, their notations, and respective values are provided in Table 3 for clarity to readers.

Data Availability

The data used in in this work are not easy to publish directly because they involve students' personal privacy, but those who want to use the model can obtain similar data from the student information published by the University's student and teaching management department or on the Internet.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (62177012, 61967005, and 62267003), Innovation Project of GUET Graduate Education (2020YCXS022 and 2021YCXS033), and Key Laboratory of Cognitive Radio and Information Processing Ministry of Education (CRKL190107).

References

[1] D. Turnbull, R. Chugh, and J. Luck, "Transitioning to E-learning during the COVID-19 pandemic: how have higher education institutions responded to the challenge?" *Education and Information Technologies*, vol. 26, no. 5, pp. 6401–6419, 2021.

[2] E. Verdú, M. J. Verdú, L. M. Regueras, J. P. de Castro, and R. García, "A genetic fuzzy expert system for automatic question classification in a competitive learning environment," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7471–7478, 2012.

[3] M. Zhu, A. R. Sari, and M. M. Lee, "A comprehensive systematic review of MOOC research: research techniques, topics, and trends from 2009 to 2019," *Educational Technology Research & Development*, vol. 68, no. 4, pp. 1685–1710, 2020.

[4] C. Romero and S. Ventura, "Educational data mining and learning analytics: an updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[5] T. Miyazoe and T. Anderson, "The interaction equivalency theorem," *The Journal of Interactive Online Learning*, vol. 9, no. 2, 2010.

[6] V. A. Romero-Zaldivar, A. Pardo, D. Burgos, and C. Delgado Kloos, "Monitoring student progress using virtual appliances: a case study," *Computers & Education*, vol. 58, no. 4, pp. 1058–1067, 2012.

[7] F. Ke, "Online interaction arrangements on quality of online interactions performed by diverse learners across disciplines," *The Internet and Higher Education*, vol. 16, pp. 14–22, 2013.

[8] N. Li, Ł. Kidziński, P. Jermann, and P. Dillenbourg, "MOOC video interaction patterns: what do they tell us?" *Design for Teaching and Learning in a Networked World*, vol. 9307, pp. 197–210, 2015.

[9] F. Zhang, D. Liu, and C. Liu, "Difficulty-based SPOC video clustering using video-watching data," *IEICE - Transactions on Info and Systems*, vol. E104.D, no. 3, pp. 430–440, 2021.

[10] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor, "Mining MOOC clickstreams: video-watching behavior vs. in-video quiz performance," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3677–3692, 2016.

[11] F. Zhang, D. Liu, and C. Liu, "Mooc video personalized classification based on cluster analysis and process mining," *Sustainability*, vol. 12, no. 7, p. 3066, 2020.

[12] J. M. J. Murre and J. Dros, "Replication and analysis of Ebbinghaus' forgetting curve," *PLoS One*, vol. 10, no. 7, p. e0120644, 2015.

[13] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.

[14] F. V. D. Sluis, J. Ginn, and T. V. D. Zee, "Explaining student behavior at scale: the influence of video complexity on student dwelling time," in *Proceedings of the third (2016) acm conference on learning@ scale*, pp. 51–60, Edinburgh, Scotland, UK, April 2016.

[15] H. Zhu, Y. Liu, F. Tian et al., "A cross-curriculum video recommendation algorithm based on a video-associated knowledge map," *IEEE Access*, vol. 6, pp. 57562–57571, 2018.

[16] Z. Kastrati, A. S. Imran, and A. Kurti, "Integrating word embeddings and document topics with deep learning in a video classification framework," *Pattern Recognition Letters*, vol. 128, pp. 85–92, 2019.

[17] E. H. Othman, S. Abdelali, and E. B. Jaber, "Education data mining: mining MOOCs videos using metadata based approach," in *Proceedings of the 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pp. 531–534, Tangier, Morocco, October 2016.

[18] S. Zhaoyu, W. Yiru, C. Pan, and Z. Huibing, "Personalized knowledge map recommendations based on interactive behavior preferences," *International Journal of Performability Engineering*, vol. 17, no. 1, pp. 36–49, 2021.

[19] Z. Shou, X. Lu, Z. Wu, H. Yuan, H. Zhang, and J. Lai, "On learning path planning algorithm based on collaborative analysis of learning behavior," *IEEE Access*, vol. 8, pp. 119863–119879, 2020.

[20] J. Zhang, X. Shao, W. Zhang, and J. Na, "Path-following control capable of reinforcing transient performances for networked mobile robots over a single curve," *IEEE Transactions on Instrumentation and Measurement*, vol. 1, 2022.

[21] D. Wang, Y. Yih, and M. Ventresca, "Improving neighbor-based collaborative filtering by using a hybrid similarity measurement," *Expert Systems with Applications*, vol. 160, p. 113651, Article ID 113651, 2020.

[22] X. Shao, Y. Shi, and W. Zhang, "Input-and-Measurement event-triggered output-feedback chattering reduction control for MEMS gyroscopes," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 9, pp. 5579–5590, 2022.

[23] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, Hong Kong, China, April 2001.

[24] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010.

[25] W. Zhang, X. Shao, W. Zhang, J. Qi, and H. Li, "Unknown input observer-based appointed-time funnel control for quadrotors," *Aerospace Science and Technology*, vol. 126, Article ID 107351, 2022.

[26] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Information Sciences*, vol. 483, pp. 53–64, 2019.

[27] X. Shao, J. Zhang, and W. Zhang, "Distributed cooperative surrounding control for mobile robots with uncertainties and aperiodic sampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18951–18961, 2022.

[28] J. Bobadilla, F. Serradilla, and J. Bernal, "A new collaborative filtering metric that improves the behavior of recommender systems," *Knowledge-Based Systems*, vol. 23, no. 6, pp. 520–528, 2010.

[29] X. Wang, L. Nie, X. Song, D. Zhang, and T. S. Chua, "Unifying virtual and physical worlds: learning toward local and global consistency," *ACM Transactions on Information Systems*, vol. 36, no. 1, pp. 1–26, 2018.

[30] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Information Sciences*, vol. 178, no. 1, pp. 37–51, 2008.

[31] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges," *ACM Computing Surveys*, vol. 47, no. 1, pp. 1–45, 2014.