information

MDPI

Article

An Effective Student Grouping and Course Recommendation Strategy Based on Big Data in Education

Yu Guo [1,2,*], Yue Chen [1], Yuanyan Xie [1] and Xiaojuan Ban [1,2]

[1] University of Science and Technology Beijing, Beijing 100083, China; chenyue@xs.ustb.edu.cn (Y.C.); yyxie@xs.ustb.edu.cn (Y.X.); banxj@ustb.edu.cn (X.B.)
[2] Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China
* Correspondence: guoyu@ustb.edu.cn

Abstract: Personalized education aims to provide cooperative and exploratory courses for students by using computer and network technology to construct a more effective cooperative learning mode, thus improving students' cooperation ability and lifelong learning ability. Based on students' interests, this paper proposes an effective student grouping strategy and group-oriented course recommendation method, comprehensively considering characteristics of students and courses both from a statistical dimension and a semantic dimension. First, this paper combines term frequency–inverse document frequency and Word2Vec to preferably extract student characteristics. Then, an improved K-means algorithm is used to divide students into different interest-based study groups. Finally, the group-oriented course recommendation method recommends appropriate and quality courses according to the similarity and expert score. Based on real data provided by junior high school students, a series of experiments are conducted to recommend proper social practical courses, which verified the feasibility and effectiveness of the proposed strategy.

Keywords: study group division; course recommendation; feature vectors; semantic analysis; clustering algorithm

check for updates

Citation: Guo, Y.; Chen, Y.; Xie, Y.; Ban, X. An Effective Student Grouping and Course Recommendation Strategy Based on Big Data in Education. *Information* **2022**, *13*, 197. https://doi.org/10.3390/info13040197

Academic Editor: Willy Susilo

Received: 20 March 2022
Accepted: 12 April 2022
Published: 14 April 2022

1. Introduction

With the rapid development of information technology and the advantages it has revealed in other fields, the relevant research on computer technology and network technology assisted teaching has gradually become the key direction of education reform. Through years of development, a tremendous amount of data in education is collected and stored and numerous novel teaching and learning methods are proposed. As a cross field of computer science and educational science, data-driven cooperative learning and course recommendation have become two attractive research directions.

Cooperative learning can improve the communication between students and create an efficient learning atmosphere, which is recognized as the most innovative and effective teaching model by front-line teachers and educators [1]. Furthermore, personalized course recommendation systems can recommend different courses for different types of students from massive educational data, thus improving students' learning efficiency and achieving continuous learning. This paper comprehensively considered the cooperative learning and the personalized course recommendation problem and designed a student grouping and course recommendation scheme based on real data to help students develop effective cooperation and build a solid foundation for lifelong learning. To achieve this goal, the below challenges need to be tackled:

(1) The commonly used feature word extraction methods are statistics-based methods and semantics-based methods. Statistics-based methods include methods using term frequency–inverse document frequency (*TF–IDF*) [2], information gain, word length and so on. Semantics-based methods include methods based on the HowNet [3]

concept and ontology. However, the above methods are not comprehensive enough in the representation of semantic information, especially in the representation of synonyms. So, a reliable feature extraction method should be proposed to accurately and comprehensively characterize students and courses.

(2) As students' grouping labels are not easy to obtain in practice, existing research mainly uses unsupervised clustering algorithms, for example, K-means, to group students. However, the traditional K-means algorithm initializes cluster centers randomly, which will lead to incorrect or uneven cluster division and cause incorrect results. Therefore, an effective student grouping strategy should be deeply studied based on the traditional K-means algorithm.

(3) Considering the goal to achieve in this paper, the item collaboration filter (ItemCF) [4] algorithm is adopted in this paper to recommend courses to the student groups. However, the traditional ItemCF algorithm has a serious cold-start problem, because it mainly uses user behaviors to calculate the similarity of items and recommend similar items. Therefore, it is impossible to recommend a new item as there is no record of it in the item-related table. Therefore, how to recommend high-quality courses that meet students' interests and solve the cold-start problem needs to be solved.

Based on the above considerations, an effective student grouping strategy and group-oriented course recommendation method, comprehensively considering characteristics of students and courses, both from the statistical dimension and the semantic dimension, are designed in this paper. The main contributions of this paper are as follows:

(1) An accurate feature extraction algorithm for representing students' characteristics is designed. This paper comprehensively selected feature words from two dimensions: First, *TF–IDF* weighting is used to select feature words from the word frequency dimension. Then, a Word2Vec [5] model is trained to select feature words from the semantic dimension. These feature words are combined to obtain the final text set representing students' characteristics to the greatest extent.

(2) An improved K-means algorithm is designed to group students. By observing the characteristics of the extracted feature words, this paper constructs a multi-dimensional vector to represent each student. Then, an appropriate grouping result is obtained by using the improved K-means algorithm, in which the method of selecting the initial cluster center is improved to ensure that the distance between the points in the cluster and the initial cluster center is less than a certain value.

(3) A group-oriented course recommendation method is ultimately proposed. This paper introduces a semantic recommendation model and expert scoring to assist in course recommendation, thus improving the quality of the recommendation results. Considering that the number of courses selected by students in each semester is generally very small compared to the number of courses, this paper uses the semantic information to solve the serious cold-start problem.

(4) A series of experiments, based on real data provided by junior high school students (12 to 15 years old), are conducted to verify the feasibility and effectiveness of the proposed strategy, which can group students of all educational levels and recommend courses (both online and offline) to them.

The rest of the paper is organized as follows: Section 2 describes related works. Section 3 gives the system framework. Section 4 introduces the details of the student grouping algorithm and course recommendation algorithm. Experiments and analyses are included in Section 5. Finally, Section 6 summarizes this paper and proposes future work.

2. Related Works

2.1. Student Grouping Strategy

Cooperative learning can not only increase the accumulation of knowledge, but also improve students' cooperation ability. So et al. explored the relationship between students' perceived level of cooperative learning and their social presence and found that there is a statistically positive correlation between them [6]. Authors in [7] found that in a group discussion, students can obtain more answers through communication and tend to communicate with different students. A strategy was designed to combine the K-means clustering algorithm and the learning styles of different students to provide a valuable reference for student grouping [8]. Tacadao et al. [9] compiled a program by using constraint logic programming to realize the student grouping. Considering the small sample size, Pang et al. used the balanced K-means algorithm to divide students into several clusters and then took samples from different groups to construct heterogeneous learning groups [10]. A K-means-based feature selection model and an AdaBoost-based behavior classification model are utilized together to achieve group online learning [11]. The above research uses mainly statistics-based methods, which may not characterize and group students properly, so a more reasonable and comprehensive student grouping strategy is very important.

2.2. Personalized Course Recommendation

The personalized recommendation system was born in the 1990s with the rise of e-commerce. Nowadays, personalized recommendation technology has been extended to many fields. As the number and types of courses increase, the amount of course-related information obtained by students also increases rapidly, so it is necessary to design course recommendation methods to address students' personal goals and interests. The course selection system is the mainstream online course platform [12]; it recommends the most suitable courses for students according to their gender, cognition, personality, learning style and performance. Aher et al. [13] used the K-means and association rule algorithm to analyze the correlation between courses and recommend them. Meson et al. used users' evaluation data to construct a course evaluation matrix to assess student performance [14]. Manouselis et al. proposed an agent-based e-learning course recommendation method that matches learners' characteristics with content attributes to recommend suitable courses for them [15]. A hybrid recommendation algorithm based on collaborative filtering to establish a course evaluation matrix in the case of sparse data, was proposed in [16]. Similarly, a content-based course recommendation method was proposed in [17]. On the condition of meeting the course requirements, Parameswaran et al. proposed the design of a recommendation system under complex constraints [18]. Si et al. proposed a big data-assisted recommendation of personalized learning resources and teaching decision support, by taking the education of spoken and written language as an example [19].

3. System Workflow

The whole strategy consists of three parts, as shown in Figure 1. The first part describes the text data preprocessing, which mainly consists of text segmentation, feature extraction and word embedding. The second part introduces the procedure of the semantic K-means clustering algorithm, in which the student feature vector is introduced to achieve effective student grouping. The third part introduces the improved ItemCF course recommendation algorithm based on similarity and expert knowledge, which can increase the ratio of the course coverage and reduce the problem of the cold-start.

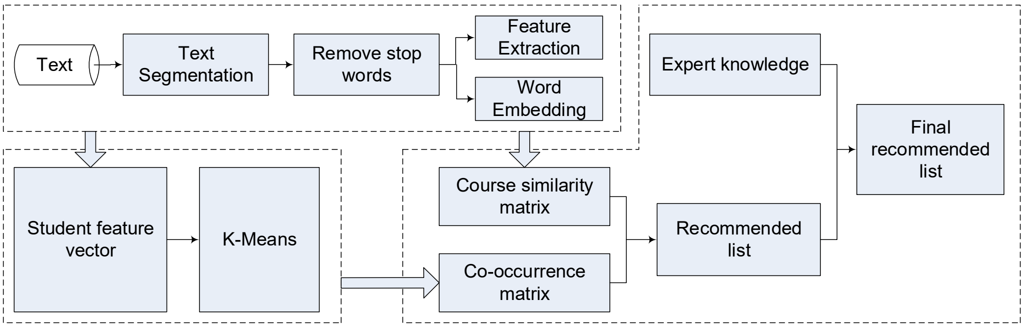


Figure 1. The workflow of the strategy proposed in this paper.

4. Student Grouping Strategy and Group-Oriented Course Recommendation Method Based on Semantics

4.1. Feature Extraction Based on TF–IDF and Word Vectors

To achieve accurate student grouping, it is necessary to reasonably represent the characteristics of the students. The weights obtained by the *TF–IDF* algorithm can make full use of the word frequency to measure the importance of words. However, this algorithm ignores the semantic relationship between words. For example, meaningless high-frequency words are often extracted, while low-frequency words with large contributions to actual meaning are ignored, which leads to the incomplete representation of students' characteristics. Therefore, this paper designs a feature extraction algorithm that combines *TF–IDF* and semantic information; it includes the following steps:

4.1.1. Text Preprocessing

Text data are a rich source of unstructured data, usually consisting of documents and presented in the form of words, sentences or paragraphs. The inherent unstructured and noisy characteristics of text data make it difficult for various machine learning and deep learning methods to deal with it directly. Therefore, it is necessary to preprocess text data before feature extraction.

The main steps of text preprocessing are as follows: (a) Establish a proper noun database and stop word thesaurus. (b) Segment the text data representing students' characteristics. This paper uses Jieba [20], the most popular Chinese word segmentation tool, to complete this step. (c) Tag the participles and types of the words, such as noun, verb, adverb or adjective. The main goal of this paper is to obtain the representation degree of students in different respects. From the perspective of word type, most representation items are composed of nouns and adjectives and, therefore, these kinds of words are screened out in this paper. (d) Screen out words with high frequency to obtain the final set of feature words describing students' interests.

The original data set D can be expressed as $D' = \{t_1, t_2, t_3, \ldots, t_n\}$ after preprocessing, where t_i is the ith word item and n is the number of words.

4.1.2. Screening Feature Words from the Statistical Dimension

One student may have multiple feature words and different students may have different preferences regarding the same interest. In this paper, the feature words in the text data are extracted from the perspectives of word frequency and semantics.

The *TF–IDF* algorithm is a mainstream feature word extraction method that uses the numbers of words in different ranges to extract the representation level of a feature word in the current text. *TF–IDF* can be divided into TF and IDF. TF is the term frequency and IDF represents the inverse document frequency. The calculation formulas of TF and IDF are shown in (1) and (2), where $n_{i,j}$ represents the number of the word t_i in a document, $\sum_k n_{k,j}$ represents the total number of occurrences of words in a document, D represents the total number of documents and $\left|\{j : t_i \in d_j\}\right|$ represents the total number of occurrences of the word-t_i in document d_j.

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}} \tag{1}$$

$$idf_i = log\frac{|D|}{\left|\{j : t_i \in d_j\}\right|} \tag{2}$$

Therefore, the calculation formula of *TF–IDF* is:

$$TF - IDF = tf_{i,j} \times idf_i = \frac{n_{ij}}{\sum_k n_{k,j}} \times log\frac{|D|}{\left|\{j : t_i \in d_j\}\right|} \tag{3}$$

The weight obtained by the *TF–IDF* algorithm can make full use of the information on word frequency to measure the importance of a word; that is, if the frequency of a word in the current text is high, its importance is high.

After the members of data set D' are arranged in descending order according to the *TF–IDF* value, the initial set of feature words $D'' = \{t_1, t_2, t_3, \ldots, t_k\}$ and the set of filter words $T = \left\{t_{k+1}, t_{k+2}, t_{k+3}, \ldots, t_{k+j}\right\}$ can be obtained, where $k < n,\ k + j = n$.

4.1.3. Screening Feature Words from the Semantic Dimension

The *TF–IDF* algorithm ignores the semantic relationships between words so that meaningless high-frequency words are often extracted while low-frequency words with large contributions to actual meaning, such as important synonymous words with low-frequencies, are ignored. This drawback leads to the incomplete representation of the text. Therefore, this paper combines the semantic information between word items with the *TF–IDF* value to obtain a better result.

Specifically, the proposed method takes the student's text data as the corpus, obtains the semantic information of each word through the Word2Vec model [11], and obtains a word vector representing the semantic information to effectively solve the problem of synonyms encountered by traditional algorithms. Using the model obtained from the data training, this paper counts the 10 closest words corresponding to the word vectors of certain words, as shown in Table 1. Through the use of synonyms, this paper can still classify students into the same group when different vocabulary expressions are used.

Table 1. Similar words based on word embeddings.

Vocabulary	Similar Words
Reading	Extracurricular reading materials, classics, storybooks, reading notes, required reading, intensive reading, bibliography, books, reading
Music	Light music, arrangement, singing, pure music, melody, electric sound, rock music, concert, piano, guitar
Unmanned aerial vehicle	Aerial photography, control, rotor, flight, aircraft, remote control, helicopter, aircraft, fixed wing

Once the word vectors are obtained, the similarity between the feature word t_i and the text data D'' can be calculated according to (4), where $embedding_{t_i}$ represents the word vector of word-t_i and $dist(\cdot)$ represents the similarity as measured by the Euclidean distance.

$$SemaSimilar(t_i, D'') = \frac{\sum_{j=1}^{m} dist\left(embedding_{t_i}, embedding_{t_j}\right)}{n} \tag{4}$$

The specific calculation process is as follows: (a) Select an undetermined feature word t_{k+1} in the set of filter words T; (b) Calculate the similarity between the undetermined word and the words in D'' according to the formula (4); (c) Repeat steps (a) and (b) until all feature words in T are processed; (d) Arrange the similar results in descending order; and (e) Screen and add the first $(m - k)$ feature words to the final set D'''.

4.2. Student Grouping Based on Semantics

4.2.1. Multidimensional Feature Vector Representation Model

Based on the characteristics of the data, this paper constructs a six-dimensional feature vector $SV = [sv_1, sv_2, sv_3, sv_4, sv_5, sv_6]$ to represent the features of the student, where the sv_i values represent ideological character, cognitive ability, sports health, talent, practical ability and interest specialty, respectively. sv_i can be calculated according to (5), where T

represents the number of feature words and its value depends on the amount of text data submitted by the student.

$$sv_i = \sum_{j=1}^{T} \frac{TF - IDF_j}{TF - IDF} Embedding_j \tag{5}$$

Then, the similarity between students $s1$ and $s2$ can be calculated according to (6), where L represents the length of the word vector.

$$dist_{s1,se} = \| sv_{s1} - sv_{s2} \|_2^2 = (\sum_{i=0}^{6} \sum_{j=0}^{L} (sv_{s1_i} - sv_{s2_i})^2)^{\frac{1}{2}} \tag{6}$$

4.2.2. Student Grouping Based on an Improved K-Means Algorithm

Based on the students' feature vectors, this paper uses the K-means algorithm to perform student grouping. The initial cluster center of the K-means is usually selected randomly, which will lead to incorrect or uneven cluster division and this cannot be ignored in subsequent course recommendations. In this paper, the method of selecting the initial cluster center is improved to ensure that the distance between the points in the cluster and the initial cluster center is less than a certain value.

The improved initial cluster center selection method is as follows: (a) The threshold value n is set, which is the number of people in a group and is the mean value of the numbers of people in groups with different K values, (b) A point C_i is randomly selected from the student data set SV and is added to the cluster center set C and removed from SV, (c) The distances between other data points and the cluster center c_i are calculated, the first n data points are selected in ascending order and they are deleted from SV, (d) If the number of data points in the set SV is less than $2n$, the members of the current set C are arranged as the initial cluster centers. Otherwise, the method returns to (b). Above steps are performed by using Algorithm 1.

Algorithm 1 Improved initial cluster center selection algorithm

```
Input: Threshold $n$; $K$; Student data set $SV$
Output: Cluster centers set $C$
$i$ = 0, $distances$ = []
while $SV$ not None do
    $num \leftarrow$ randint(0, len($SV$))
    $C_i \leftarrow SV[num]$
    remove $C_i$ from $SV$
    for each $x \in SV$ do
        $distances[x]$ = dist($C_i$, $x$)
    end for
    select first $n$ points from sort(distances) and remove them from $SV$
    if len($SV$) < $2n$
        $C_{i+1}$ = center($SV$)
        break
    end if
    $i$ += 1
end while
```

Under the constraints of Algorithm 1, this paper uses the traditional K-means algorithm to obtain student grouping results. The basic process of K-means is as follows: (a) Take C as the initial cluster centers; (b) For each sample x_i in the data set, calculate its distance to all points in C and divide it into the class with the smallest distance; (c) For each group N_i, recalculate its cluster center C_i' using (7), that is, the centroid of all samples belonging to this group; and (d) Repeat (b) and (c) until the termination conditions are met. The student grouping method is performed by using Algorithm 2.

$$C_i' = \frac{1}{n} \sum_{x \in N_i} x \tag{7}$$

Algorithm 2 Student grouping method based on an improved K-means algorithm

Input: Student feature and *TF–IDF* value; *K* in range(*M*); Tolerance ε
Output: M kinds of clusters about K values

```
for each $K$ do
    $C \leftarrow$ Select $K$ cluster centers by Algorithm 1
    $E \leftarrow \infty$
    while $E > \varepsilon$ do
        for each $sv \in SV$ do
            for each $c \in C$ do
                $sv \in$ mindist(cluster($c$))
            end for
        $E \leftarrow 0$
        for each $c \in C$ do
            for each $sv \in$ cluster($c$) do
                $E$ += dist($sv$, $c$)
            end for
            update $c$
        end for
    end while
end for
```

4.3. Personalized Course Recommendation Scheme Based on Student Grouping

This paper proposes a hybrid course recommendation scheme based on student grouping, which consists of the following steps: (a) Based on the interest characteristics of students, the historical records of student groups and the ItemCF algorithm are used to perform course recommendation; (b) The similarity between classes is used to solve the cold-start problem of the item recommendation system and to increase the novelty of the course recommendation results; and (c) Expert knowledge is added to further improve the quality of the recommendation results.

In this paper, a student group is considered a single user, so the first step is to convert the features of a group into the features of a single user, that is, to normalize the data. Each student in the group has different numbers and categories of selected courses. To normalize the data, this paper designs a record matrix *CoM*, as shown in (8):

$$CoM = [G_1, G_2, \ldots, G_m] = \begin{bmatrix} \frac{r_{1,1}}{m_1} & \frac{r_{2,1}}{m_2} & \cdots & \frac{r_{num,1}}{m_1} \\ \frac{r_{1,2}}{m_1} & \frac{r_{2,2}}{m_2} & \cdots & \frac{r_{num,2}}{m_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{r_{1,n}}{m_1} & \frac{r_{2,n}}{m_2} & \cdots & \frac{r_{num,n}}{m_{num}} \end{bmatrix} \tag{8}$$

where G_i represents the course co-occurrence vector of the *i*th student group, m is the number of student groups, n is the number of courses, $r_{i,j}$ represents the total number of students in the *i*th group who have chosen course *j* before and m_i is the number of students in group *i*.

After normalization, the ItemCF algorithm is used to perform the course recommendation, which is divided into two main steps: (a) Calculating the similarity between courses; and (b) Generating a recommendation list for student groups according to the similarity of courses and students' historical behaviors.

ItemCF uses (9) to define the similarity between items i and j, where $N(i)$ is the number of students who like course i and $N(j)$ is the number of students who like course j. Two courses can be determined to be similar because they are commonly enjoyed by many students; i.e., the students can "contribute" to the judgement of course similarity through their historical interest lists.

$$\omega_{i,j} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i) \cap N(j)|}} \tag{9}$$

Although the traditional ItemCF algorithm can achieve relatively good recommendation results according to students' interests, it has a low ratio of course coverage and hinders students from broadening their vision. This paper considers using semantic information to increase the ratio of course coverage and reduce the cold-start problem.

This paper uses the process of obtaining the course representation vector that is described in Section 4.1. Then the cosine similarity calculation method is used to measure the semantic similarity between courses; its value is between -1 and 1, as shown in (10). The greater this value is, the greater the similarity. Here, cv_u is the feature vector of course u and cv_v is the feature vector of course v. $I(u)$ is a nonzero element in cv_u and $cv_{u,i}$ represents the ith nonzero element in cv_u.

$$cos_{u,v} = \frac{cv_u \cdot cv_v}{|cv_u| \cdot |cv_v|} = \frac{\sum_{i \in I(u) \cap I(v)} cv_{u,i}, cv_{v,i}}{\sqrt{\sum_{i \in I(u)} cv_{u,i}^2} \sqrt{\sum_{i \in I(v)} cv_{v,i}^2}} \tag{10}$$

Considering that the number of courses selected by students in each semester is generally very small compared to the number of courses, there will be a large number of similar courses with zero similarity in the similarity matrix, which is a serious cold-start problem. Therefore, this paper uses the semantic information calculated by (10) together with the original similarity calculated by (9) to redefine the similarity between courses i and j, as shown in (11), where α_0 and β_0 are normalization factors that can be dynamically adjusted according to specific needs.

$$\omega'_{i,j} = \alpha_0 \omega_{i,j} + \beta_0 cos_{i,j} = \alpha_0 \frac{|N(i) \cap N(j)|}{\sqrt{|N(i) \cap N(j)|}} + \beta_0 \frac{cv_i \cdot cv_j}{|cv_i| \cdot |cv_j|} \tag{11}$$

Considering that courses are organized by institutions with different resources and qualifications, the quality of the courses should also be a metric. In this paper, expert knowledge is also added as a penalty, which is a comprehensive score obtained from a review by relevant experts regarding different dimensions. Therefore, a student's interest in a course can be calculated by (12), which indicates that the more similar a course is to the preferences and habits in a student's records, the more likely it is to obtain the top-ranking position in the student's recommendation results. In Equation (12), α_1 and β_1 are normalization factors; $S(j, K)$ represents the K most similar courses to course j; $G_{u,i}$ represents the interest of student group u in course i, which corresponds to the value in row u and column i in (8); and es_j is the score obtained from experts for course j.

$$p_{u,j} = \alpha_1 \sum_{i \in N(u) \cap S(j,K)} \omega'_{i,j} G_{u,i} + \beta_0 \cdot es_j \tag{12}$$

The course recommendation scheme is performed by using Algorithm 3.

Algorithm 3 Personalized Course Recommendation Scheme

```
Input: Train Set $t$; Usergroup_id $u$; Similarity Matrix $W$; $N$
Ouput: Recommendation results $Results$
Initial $Results \leftarrow$ dict
Initial $ru \leftarrow t[u]$
for each $i$, $p_{u,i} \in ru.items()$ do
    $kl \leftarrow$ sorted($W[i].items()$, key=itemgetter(1), reverse=True)[0:N]
    for each $j$, $wj \in kl$ do
        if $i == j$ do
            continue
        end if
        $Results[j] \mathrel{+}= p_{u,i} * wj$
    end for
end for
```

5. Experimental Results and Analysis

5.1. Data Set

The data set used in this paper consists of comprehensive quality evaluation data and open social practice course data from 103 junior middle schools in the Fengtai District and Xicheng District in Beijing, which is provided by the Beijing Municipal Education Commission. The details of the data set are shown in Table 2.

Table 2. Experiment-related data sets.

Data Set	Information	Example	Data Set Size
Student information	Student ID	10037952	27,275 items
	School name	Beijing No. 66 High School	
Comprehensive quality evaluation data	Student ID/Activity type	10037952/Practice	270 M
	Activity record	I participate in the open science practice—"Playing with the Solar system". I learned that the center of the solar system is the sun and I also learned about the eight planets.	
Open social practice course data	Course ID/Course name	10013-c2/Centrifugal force in life	1018 items
	Expert score	64	
	Course record	By understanding the structure of the dryer, students can learn the relevant scientific knowledge of centrifugal movement.	
Course selection records	Student ID/Course ID	10037952/636	133,047 items

The data are mainly provided by students themselves according to their personal interests, so they can fully reflect the personal characteristics of the students and can support the study of student clustering and the course recommendation algorithm proposed in this paper.

5.2. Evaluation Metrics

5.2.1. Evaluation Metrics for the Student Grouping Method

To verify the performance of the student clustering method, this paper selects 1000 students for the experiment and measures performance in two aspects. First, this paper verifies the effect of the improved initial cluster selection scheme from the convergence effect. Second, this paper evaluates the clustering results with two common evaluation indices: the intra-class compactness (*CP*) and the Davies–Bouldin index (*DBI*). The calculation method of *CP* is shown in (13) and (14):

$$\overline{CP_i} = \frac{1}{|\Omega_i|} \sum_{x_i \in \Omega_i} \parallel x_j - C_i \parallel \quad (13)$$

$$\overline{CP} = \frac{1}{K} \sum_{i=1}^{K} \overline{CP_i} \quad (14)$$

where $\overline{CP_i}$ represents the average distance between each point and the cluster center in one group, K represents the number of clusters and C_i represents the center of cluster i. The lower the CP is, the smaller the clustering distance and the more similar the members are.

The calculation method of *DBI* is shown in (15) and (16):

$$DBI = \frac{1}{k} \sum_{I=1}^{K} max_{j \neq i} \left(\frac{\overline{S_i} + \overline{S_j}}{\parallel C_i - C_j \parallel_2} \right) \quad (15)$$

$$\overline{S_i} = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} (x_j - C_i)^2 \right)^{\frac{1}{2}} \quad (16)$$

K is the number of groups, T_i is the number of students in cluster i, C_i represents the centroid of cluster i and $\overline{S_i}$ represents the mean square error of cluster i. The smaller the *DBI* value is, the better the performance of the clustering algorithm.

In addition, to obtain a more balanced grouping result, this paper uses the unevenness indicator, as shown in (17):

$$unevenness = \frac{1}{N} sum_{i=1}^{K} \left| T_i - \frac{N}{K} \right| \quad (17)$$

here, N represents the total number of students. The smaller the unevenness is, the more normal the number distribution will be. Therefore, the comprehensive indicator to determine the best K is defined by (18), where α and β are normalization factors that are set to 0.7 and 0.3, respectively, in this paper according to many experiments. The goal of the proposed student grouping method is to find a suitable K with a relatively small comprehensive indicator (*CI*).

$$CI = \alpha \times DBI + \beta \times unevenness \quad (18)$$

Moreover, this paper adopts a manual evaluation to measure the effectiveness of the student grouping method. If a student's feature words are totally different from those of other students in the group, it is assumed that he or she is grouped incorrectly. Formula (19) defines the grouping accuracy rate, where n_i represents the number of students in group i and m_i represents the number of students who are grouped incorrectly.

$$SGC = \frac{1}{N} \sum_{i=1}^{N} \frac{n_i - m_i}{n_i} \quad (19)$$

5.2.2. Evaluation Metrics for the Course Recommendation Algorithm

Since this paper uses the top-N recommendation method to obtain the course recommendation list, the precision and recall rate are used to measure the performance of the recommendation algorithm. The precision is defined by (20), where $R(u)$ is the recommendation list given to the user and $T(u)$ is the course list chosen by the user on the given test data.

$$Precision = \frac{\sum_{u \in u} |R(u) \cap T(u)|}{sum_{u \in U} |R(u)|} \quad (20)$$

The recall is defined by (21):

$$Recall = \frac{sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in u}|T(u)|} \tag{21}$$

To comprehensively evaluate the precision and recall, this paper tests the algorithm on different N to obtain a group of precision and recall values. In addition, this paper uses the average popularity and average score of the recommended courses to verify that the recommendation system can effectively solve the cold-start problem and recommend a high-quality course, as shown in (22), where $p(u)$ represents the total number of times course u appears in all users' recommendation lists. Low popularity proves that the cold-start problem is effectively solved.

$$popularity = \frac{\sum_{u\in u}\log(1+p(u))}{\sum_{u\in u}|R(u)|} \tag{22}$$

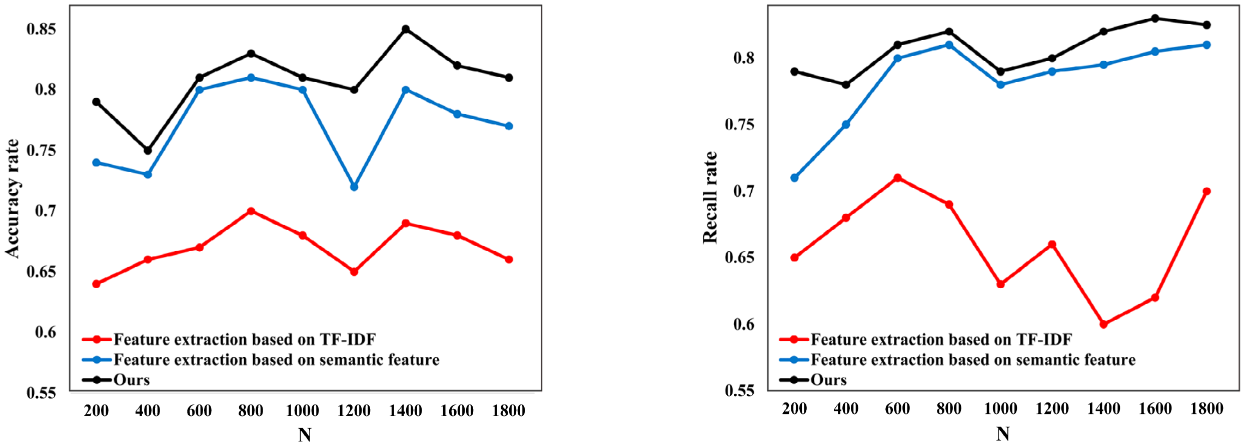
The average score of a recommended course is defined by (23):

$$RAS = \frac{\sum_{u\in u} es_u}{\sum_{u\in u}|R(u)|} \tag{23}$$

5.3. Experimental Results

5.3.1. Verification of the Feature Extraction Method

To verify the text representation effect of the feature extraction algorithm, this paper compares the classification effect of different feature extraction algorithms on the same classification model. Figure 2a,b shows the accuracy and recall comparisons of different feature extraction algorithms. It can be seen from Figure 2 that the algorithm proposed in this paper can achieve better results when dealing with different numbers of text data sets, so it has a better effect on the representation of text information and can extract low-frequency information with rich semantics.

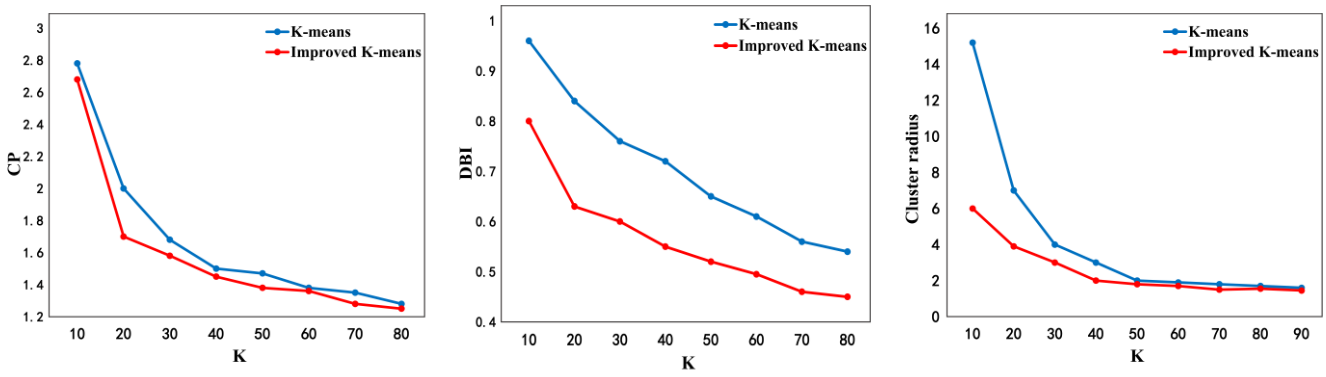


(**a**) (**b**)

Figure 2. Comparison of different feature extraction algorithms: (**a**) Accuracy comparison; (**b**) Recall comparison.

5.3.2. Verification of the Student Grouping Method

To verify the performance of the improved K-means student grouping method, the improved K-means algorithm was used to divide 1000 students into K groups to compare the performance according to *CP*, *DBI* and *Cluster Radius*, as shown in Figure 3. The above metrics of the improved K-means algorithm are all smaller than those of the traditional one, which verifies that the improved student grouping method has a better performance.



(**a**) (**b**) (**c**)

Figure 3. *CP*, *DBI* and *Cluster Radius* comparison results: (**a**) *CP* comparison; (**b**) *DBI* comparison; (**c**) *Cluster Radius* comparison.

To prove that the improvements of the K-means algorithm are statistically significant, this paper adopted the "Mann-Whitney U test" [21], one of the non-parametric statistical tests, to provide statistical evidence. As the goal of the proposed student grouping method is to find a suitable K with a relatively small CI, this paper conducted a series of experiments to record the smallest CI values of the improved K-means algorithm and the original K-means algorithm, based on data from same students. A total of 15 sets of CI values for different students are finally obtained, as shown in Table 3. The H_0 hypothesis is "CI values of the original K-means algorithm $\leq$ CI values of the improved K-means algorithm". The final p value is 0.004203, which is smaller than the significance level α (0.05). Therefore, H_0 is rejected, which means that the randomly selected CI value of the original K-means is considered to be greater than the randomly selected CI value of the improved K-means. Hence, the improvements in the improved K-means are statistically significant.

Table 3. Statistical test on the improved K-means algorithm and the original one.

Algorithms	**Significance Level (α):**	p	**No. of Samples**														
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Traditional Algorithm	0.05	0.004203	1.34	1.88	1.86	1.51	1.56	1.48	2.31	1.33	1.29	1.39	1.21	1.49	1.90	1.61	2.13
Improved Algorithm			0.98	1.60	1.43	1.21	1.33	1.23	1.89	0.93	0.96	0.99	0.96	1.34	1.39	1.26	1.69

The practical effect of the proposed student grouping method is also verified, as shown in Figure 4. Typical student data from four schools were selected as the experimental data: Beijing Dacheng School (school_1), Beijing No. 66 High School (school_2), Beijing No. 18 High School (school_3) and Beijing No. 8 High School (school_4), with 123, 456, 688 and 1256 students and a total of 104,739 comprehensive quality evaluation data.

The suitable clustering numbers K of the four schools are determined according to (18). Taking school_1 and school_2 as examples, suitable grouping numbers are $K = 13$ and $K = 25$. It can be seen from Figure 4 that, with the increase in K, DBI will gradually decrease and the imbalance of grouping will first decrease and then increase. Therefore, for each school, there is a suitable K. From the trend of the data, the student grouping method proposed in this paper can obtain groups for different numbers of students, which shows good generalization. Similarly, suitable K values of school_3 and school_4 are 40 and 61, respectively.

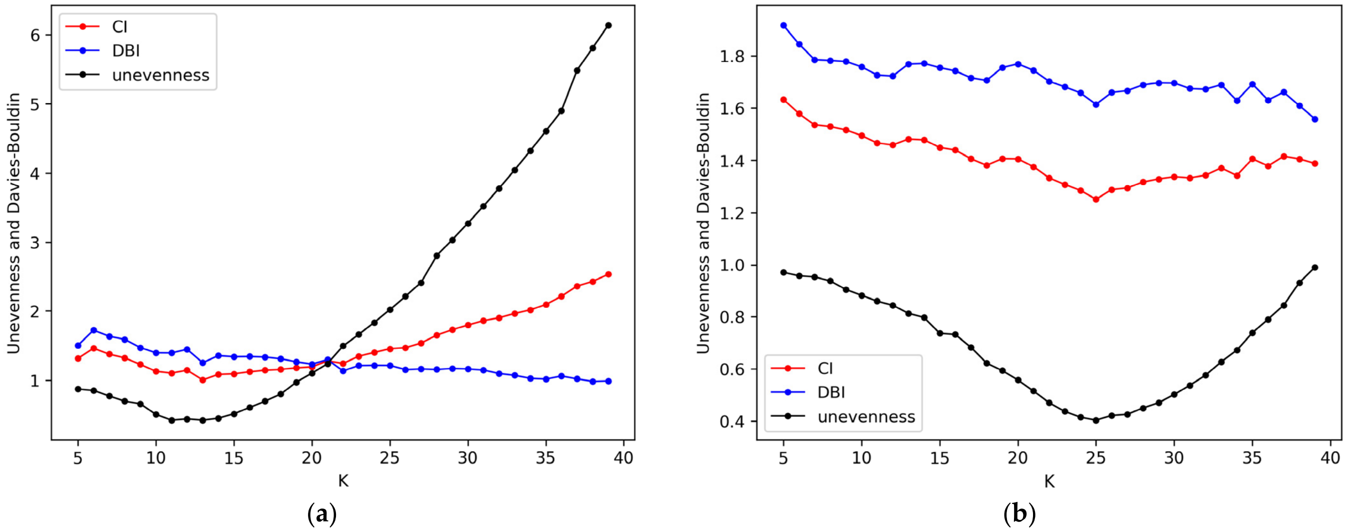


Figure 4. Determination of the suitable number of groups: (**a**) Determining the suitable *K* for school_1; (**b**) Determining the suitable *K* for school_2.

After determining the suitable *K* for each school, this paper presents the grouping results intuitively by dimensionality reduction; as shown in Figure 5, the axes represent the features obtained after dimensionality reduction. The data points for each group (distinguished by different colors) show a tendency of centralized distribution around the cluster center. Students with similar characteristics can be divided into the same cluster and, at the same time, the distribution of the number of students in each cluster is also relatively normal. The clustering result directly reflects that the student grouping method proposed in this paper has good performance in the student grouping task and enables grouping-oriented course recommendation.

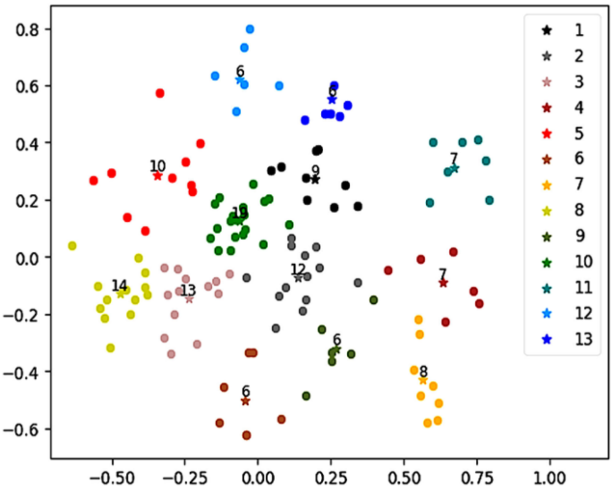


Figure 5. Grouping results of students in school_1. The axes represent the features obtained after dimensionality reduction. The data points for each group are marked by different colors.

Table 4 shows the accuracy of the interest group classification results according to (19) with the suitable *K*. It can be seen from Table 4 that the accuracy of the improved clustering algorithm for each school is above 90%. Combined with the student grouping results shown above, it can be verified that the K-means algorithm based on semantics proposed in this paper can effectively group students based on their characteristics.

Table 4. The accuracy of the clustering results of each school.

School	School_1	School_2	School_3	School_4
Accuracy of clustering	93.2%	90.5%	91.6%	90.2%

5.3.3. Verification of the Course Recommendation Method

Table 5 shows a performance comparison between the hybrid recommendation algorithm proposed in this paper and the traditional recommendation algorithm. This paper takes the course selection data of 1818 groups of students from 103 schools and divides them into training data and test data for the course recommendation experiment. The performance of algorithm precision, recall and popularity is evaluated under different recommendation numbers *N*. The above metrics are tested using 10-fold cross validation to obtain mean values.

Table 5. Performance comparison between the improved recommendation algorithm and the traditional recommendation algorithm.

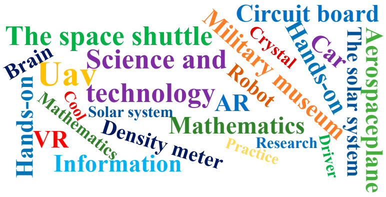
Algorithms	Metrics	Average	Number of Recommendation Results								
			2	3	4	5	6	7	8	9	10
Traditional Algorithm	Precision	0.69	0.73	0.76	0.78	0.76	0.72	0.68	0.64	0.61	0.53
	Recall	0.65	0.66	0.71	0.72	0.72	0.68	0.65	0.61	0.59	0.55
	Popularity	0.77	0.77	0.79	0.82	0.81	0.79	0.76	0.73	0.72	0.71
Improved Algorithm	Precision	0.63	0.70	0.74	0.74	0.71	0.67	0.62	0.55	0.51	0.50
	Recall	0.60	0.62	0.67	0.68	0.68	0.61	0.59	0.55	0.52	0.48
	Popularity	0.31	0.31	0.32	0.31	0.30	0.31	0.30	0.31	0.32	0.30

Compared with the traditional algorithm, the semantic and expert knowledge-based recommendation system proposed in this paper has a slight decline in precision and recall. This is because when the course similarity matrix is added to the traditional recommendation algorithm, many "new courses" that have not been selected before are added to the recommendation system. When a new course is similar to a previous one, the system will recommend it, thus causing a decline in precision and recall. However, the mean value for popularity in the hybrid recommendation system proposed in this paper is 0.31 (when a course is selected by many students, the course is very popular), which is significantly lower than that in the traditional method, indicating that the recommended courses are new courses or ones that few students have chosen before. Therefore, the hybrid recommendation algorithm proposed in this paper can effectively solve the cold-start problem.

Table 6 shows the list of courses recommended by the course recommendation system proposed in this paper and part of the word cloud of specific words for each student group. The relationship between students' characteristics and courses can be intuitively seen from the cloud of characteristic words. The courses recommended to the first group of students are mainly in the high-technology direction, such as "UAV" and "rocket". This kind of association between the feature word cloud and the list of recommended courses is also found in the second group.

Table 7 shows the indicators of the course recommendation results for each school. The precision, recall, popularity and course quality indicators of the course recommendation results remain at the same levels, which indicates that the course recommendation system proposed in this paper is stable. Table 7 shows that the recommendation algorithm proposed in this paper can recommend courses based on students' characteristics. At the same time, as the accuracy of the student grouping method has a direct impact on the recommendation results, it proves the accuracy of the overall student grouping and course recommendation strategy.

Table 6. Examples of feature words cloud and course recommendation results.

Word Cloud of Specific Words	Recommended Courses
	Exploration of 3D holographic projection
	Human body sensor car
	UAV flight principles and aerial photography experience
	From recording metal to memory metal
	Overview of rockets
	Luban No. 7
	Small world—Leeuwenhoek microscope
	Hydraulic mechanical arm production
	Mortise and tenon chair
	Homemade remote-control vehicle

Circuit board
The space shuttle
Military museum
Crystal
Hands-on
Car
The solar system
Aerospaceplane
Brain
Uav
Science and
Robot
technology
AR
Cool
Mathematics
Solar system
Mathematics
Density meter
Research
Practice
Driver
Hands-on
VR
Information

Widen the vision
Mechanical arm
Fan
Jigsaw puzzle
The solar energy
Terrain
Hydraulic press
Humidifier
Invention
Practice
Cooperation
Activity
Physical
Plant
Print
Homemade
Wind
Building
Quality
Achievement
Remote control car
Landscape

Table 7. Indicators of the course recommendation results for each school.

Data Set	Average Precision	Average Recall Rate	Average Popularity	Average Course Score
School_1	0.65	0.61	0.36	0.70
School_2	0.69	0.68	0.33	0.68
School_3	0.61	0.57	0.31	0.64
School_4	0.58	0.55	0.31	0.62

6. Conclusions

The concept of quality education has been a research hotspot in the field of education since its introduction. Based on students' interests, this paper comprehensively considered the characteristics of students and courses both from a statistical dimension and a semantic dimension, and the proposed effective student grouping strategy and group-oriented course recommendation method, thus improving students' cooperation ability and lifelong learning ability. The advantages of the proposed strategy are proven through detailed experiments, which also gave a demonstration of grouping students of all educational levels and recommending courses to them.

In future work, we will focus on combining semantic analysis with emotion analysis to construct more detailed user feature portraits and update students' feature models considering time factors.

Author Contributions: Conceptualization, Y.G., Y.C., Y.X. and X.B.; formal analysis, Y.G., Y.C. and Y.X.; methodology, Y.G.; project administration, Y.G. and Y.X.; software, Y.G. and Y.C.; supervision, X.B.; writing—original draft, Y.G.; writing—review and editing, X.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the China Postdoctoral Science Foundation under grant 2020M680352, partially funded by the Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB under grant 2020BH011, partially funded by the National Natural Science Foundation of China under grant 61772068, and partially funded by the Fundamental Research Funds for the Central Universities under grant FRF-TP-20-063A1Z.

Data Availability Statement: Related codes of this study are on GitHub: https://github.com/fovyu/Student_Grouping_and_Course_Recommendation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sofroniou, A.; Poutos, K. Investigating the Effectiveness of Group Work in Mathematics. *Educ. Sci.* **2016**, *6*, 30. [CrossRef]
2. Martineau, J.; Finin, T. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In Proceedings of the 2009 3rd AAAI International Conference on Weblogs and Social Media (ICWSM), San Jose, CA, USA, 17–20 May 2009. Available online: https://ojs.aaai.org/index.php/ICWSM/article/view/13979/13828 (accessed on 18 October 2021).
3. Dong, Z.; Qiang, D.; Hao, C. HowNet and its computation of meaning. In Proceedings of the 2010 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010.
4. Linden, G.; Smith, B.; York, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* **2003**, *7*, 76–80. [CrossRef]
5. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the 2013 1st International Conference on Learning Representations (ICLR), Scottsdale, AZ, USA, 2–4 May 2013.
6. So, H.; Brush, T. Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. *Comput. Educ.* **2008**, *51*, 318–336. [CrossRef]
7. Ruane, R. A Study of Student Interaction in an Online Learning Environment Specially Crafted for Cross-Level Peer Mentoring. Ph.D. Thesis, Philadelphia, PA, USA, 2012.
8. Liu, Q.; Ba, S.; Huang, J.; Wu, L.; Lao, C. A study on grouping strategy of collaborative learning based on clustering algorithm. In Proceedings of the International Conference on Blended Learning, Hong Kong, China, 27–29 June 2017; Springer: Cham, Switzerlnad, 2017; pp. 284–294. [CrossRef]
9. Tacadao, G.; Toledo, R.P. Forming Student Groups with Student Preferences Using Constraint Logic Programming. In Proceedings of the 2016 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), Varna, Bulgaria, 7–10 September 2016. [CrossRef]
10. Pang, Y.; Xiao, F.; Wang, H.; Xue, X. A Clustering-Based Grouping Model for Enhancing Collaborative Learning. In Proceedings of the 2014 13th International Conference on Machine Learning and Applications (ICMLA), Detroit, MI, USA, 3–6 December 2014. [CrossRef]
11. Zhu, T.B.; Wang, L.; Wang, D. Features of Group Online Learning Behaviors Based on Data Mining. *Int. J. Emerg. Technol. Learn.* **2022**, *17*, 34–47. [CrossRef]
12. Wang, Y.H.; Tseng, M.H.; Liao, H.C. Data mining for adaptive learning sequence in English language instruction. *Expert Syst. Appl.* **2009**, *36*, 7681–7686. [CrossRef]
13. Aher, S.B.; Lobo, L.M.R.J. Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowl. Based Syst.* **2013**, *51*, 1–14. [CrossRef]
14. Meson, G.; Dragovich, J. Program assessment and evaluation using student grades obtained on outcome-related course learning objectives. *J. Prof. Issues Eng. Educ. Pract.* **2010**, *27*, 1315–1318. [CrossRef]
15. Manouselis, N.; Sampson, D. Agent-Based E-Learning Course Recommendation: Matching Learner Characteristics with Content Attributes. *Int. J. Comput. Appl.* **2003**, *25*, 50–64. [CrossRef]
16. Xiao, J.; Wang, M.; Jiang, B.; Li, J. A personalized recommendation system with combinational algorithm for online learning. *J. Ambient Intell. Humanized Comput.* **2018**, *9*, 667–677. [CrossRef]
17. Wang, H.; Zhang, P.; Lu, T.; Gu, H.; Gu, N. Hybrid recommendation model based on incremental collaborative filtering and content-based algorithms. In Proceedings of the IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD), Wellington, New Zealand, 26–28 April 2017; pp. 337–342. [CrossRef]
18. Parameswaran, A.; Venetis, P.; Garcia-Molina, H. Recommendation systems with complex constraints. *ACM Trans. Inf. Syst.* **2011**, *29*, 1–33. [CrossRef]
19. Si, H.J. Big Data-Assisted Recommendation of Personalized Learning Resources and Teaching Decision Support. *Int. J. Emerg. Technol. Learn.* **2022**, *17*, 19–32. [CrossRef]
20. Jieba Chinese Word Segmentation Tool. Available online: https://github.com/fxsjy/jieba. (accessed on 12 November 2021).
21. Prospere, K.; McLaren, K.; Wilson, B. Plant Species Discrimination in a Tropical Wetland Using in Situ Hyperspectral Data. *Remote Sens.* **2014**, *6*, 8494–8523. [CrossRef]