Research on student portrait system based on canopy and k-means algorithm

S Z Zhang[#a], M Z Shen[*b], Y R Yu[b]

[a]School of Software, Zhengzhou University of Light Industry, Zhengzhou, Henan, China; [b]School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, China

ABSTRACT

Aiming at the problems of incomplete information and low data mining efficiency in the current student management system, a college student portrait system is established based on Hadoop big data processing technology. The system collects student data from various business platforms in colleges and universities, and uses HDFS for data storage; uses canopy and k-means based clustering algorithms for multi-dimensional analysis of student data; uses Echart tool to visualize the analysis results and generate student portraits. Experiments show that the student portrait system based on canopy and k-means can describe students' images in multiple dimensions and help schools understand students more comprehensively.

Keywords: Student portrait, Smart campus, Data analysis, Canopy, K-means, Big data

1. INTRODUCTION

With the construction of digital campuses and smart campuses, various intelligent application systems in colleges and universities have accumulated a large amount of student data[1-3]. How to use these data to construct multi-dimensional student data portraits and help school administrators to classify and manage students more scientifically is the goal of smart campus construction[4].

At present, the research on campus big data has made significant progress[5-7]. Yin aggregates and analyzes the data generated during students' learning, and builds student portrait tag libraries of different dimensions based on the results. Use longitudinal labels to perform variance analysis on statistical data to understand the relationship between each label and grades, and to help students learn professional courses more efficiently[8]. Perikos uses the decision tree mining method to extract rules based on the student’s usual evaluation score data to predict the student’s course performance, help teachers understand the student’s learning situation, and improve the quality of teaching[9]. Liu used the data mining K-means clustering method to analyze a total of 23.843 million online data of 3,245 students of a certain grade in a university in 4 years, and constructed a student network portrait[10].Based on personalized multiple regression and matrix decomposition methods, Elbadrawy accurately predicts student performance in future courses and classroom assessments, helping students choose their own majors and courses[11]. Ge Suhui obtained the dynamic behavior trajectory data of students on the campus through various intelligent terminals in the smart campus, performed cluster analysis on the data, and generated student portraits based on the feature matrix, so that university management agencies and teachers could grasp the living conditions of students[12].

It can be seen from the existing research that the research dimension of campus big data is not comprehensive enough. Aiming at the current problems, according to the student data in various information management platforms of colleges and universities, a Hadoop-based college student portrait system is established. By designing a student behavior description index system, using machine learning-related algorithms to mine and analyze student behavior data, and construct student data portraits from multiple dimensions, it helps school administrators manage students in a personalized manner.

#zhsuzhi@zzuli.edu.cn; *758828804@qq.com

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122601W · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637367

2. SYSTEM DESIGN

The student portrait system architecture is divided into data acquisition and processing layer, data storage layer, data analysis layer and visual display layer. First, build an experimental environment based on the Hadoop big data framework. Since the student data comes from various intelligent application systems in the university, it is necessary to use the sqoop data exchange tool to synchronously integrate the student data in the school's Oracle database into the Hive data warehouse, and the data files are stored in the HDFS distributed system. Mining and analysis of student behavior data using improved k-means algorithm. Finally, use Java Web technology and ECharts to visualize the analysis results and provide friendly user interaction functions. The system workflow is shown in Figure 1.

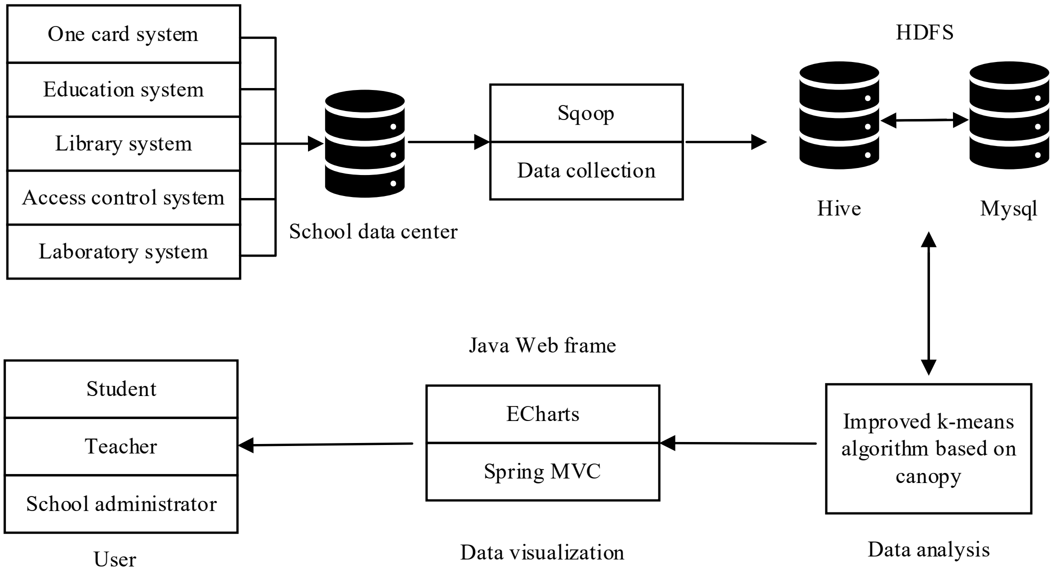

Figure 1. Student data portrait system workflow.

3. ANALYSIS METHODS

3.1 Principle of k-means algorithm

K-means is an iteratively solved cluster analysis algorithm, where k refers to the number of clusters and means is understood as the mean of the data in each class. K-means clustering is one of the most popular and simple clustering algorithms, and is still widely used in many fields today due to its simplicity, efficiency, and ease of implementation.

The basic idea of the k-means algorithm is that for a given sample set X=$\{x_1,x_2,\dots,x_m\}$ the sample set is divided into k clusters according to the distance d between samples. The calculation method of the least square error of the cluster C=$\{C_1,C_2,\dots,C_k\}$ obtained by the k-means algorithm is as follows.

$$E=\sum_{i=1}^{k}\sum_{x\in C_i} x-\mu_{i2}^{2} \tag{1}$$

$$u_i=\frac{1}{\left|C_i\right|}\sum_{x\in C_i} x \tag{2}$$

The distance calculation between sample objects is generally measured by Euclidean distance. The calculation method of Euclidean distance is as follows.

$$d(x,y)=\sqrt{\sum_{i=1}^{m}\left(x_i-y_i\right)} \tag{3}$$

The flow chart of the k-means algorithm is shown in Figure 2.

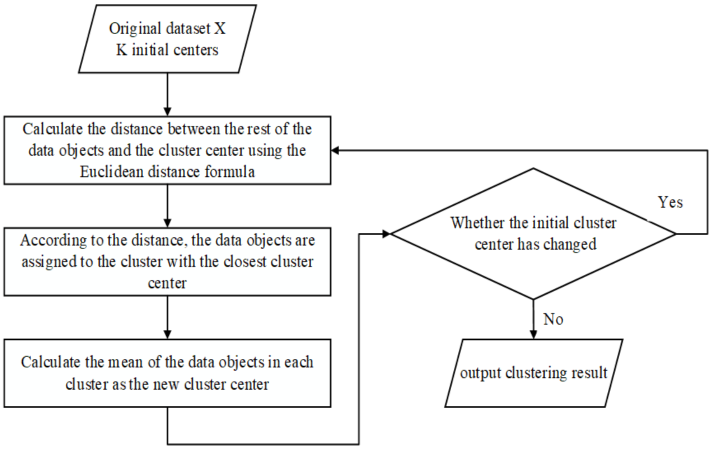
Original dataset X
K initial centers
Calculate the distance between the rest of the data objects and the cluster center using the Euclidean distance formula
According to the distance, the data objects are assigned to the cluster with the closest cluster center
Calculate the mean of the data objects in each cluster as the new cluster center
Whether the initial cluster center has changed
Yes
No
output clustering result

Figure 2. Flow chart of k-means algorithm.

3.2 Improved k-means algorithm

The k value of the traditional k-means algorithm needs to be preset, and the selection of the initial cluster center is random, so the result of the algorithm changes with the selection of the center point, which may lead to a local optimum. The canopy algorithm is a fast clustering technique. Although the accuracy is low, it cannot give accurate cluster results, but it can give the optimal number of clusters. You can use canopy clustering to first perform coarse clustering on the data, and use the number of clusters and cluster centers of the canopy algorithm as the input parameters of the k-means algorithm to complete the fine clustering of the data set. The process of using the canopy algorithm to help the k-means algorithm to select the initial cluster center is shown in Figure 3.

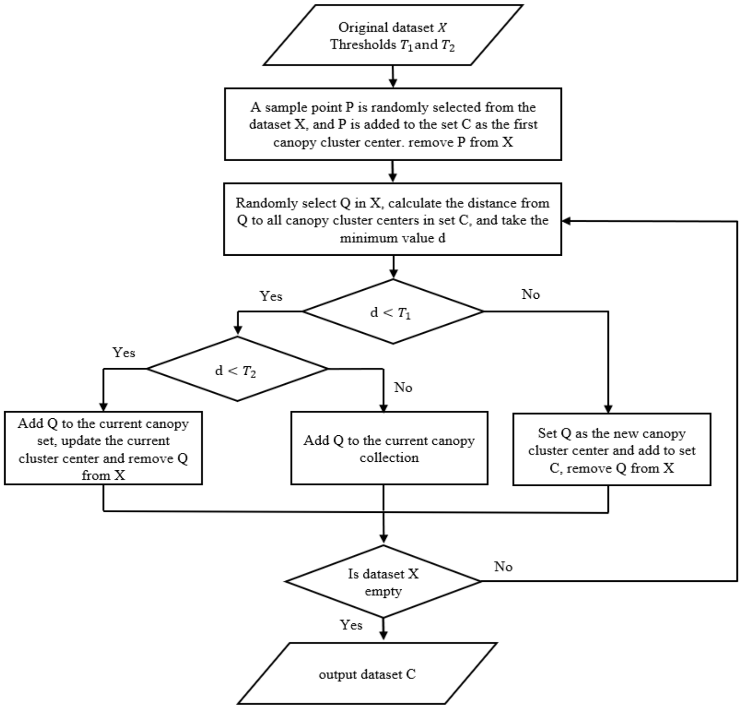
Original dataset X
Thresholds T_1 and T_2
A sample point P is randomly selected from the dataset X, and P is added to the set C as the first canopy cluster center. remove P from X
Randomly select Q in X, calculate the distance from Q to all canopy cluster centers in set C, and take the minimum value d
$d < T_1$
Yes
No
$d < T_2$
Yes
No
Add Q to the current canopy set, update the current cluster center and remove Q from X
Add Q to the current canopy collection
Set Q as the new canopy cluster center and add to set C, remove Q from X
Is dataset X empty
No
Yes
output dataset C

Figure 3. Canopy algorithm to determine cluster centers.

The specific algorithm is as follows:

Input: dataset $X = \{x_1, x_2, \ldots, x_n\}$

Step 1: Determine two distance thresholds T_1 and T_2 through cross-validation tuning or prior knowledge, where $T_1 > T_2$.

Step 2: Randomly select a sample point P from the dataset X as the first canopy cluster center, and add P to the set C. Remove P from L.

Step 3: Randomly select a sample Q from set X, and calculate the Euclidean distance from Q to each canopy cluster center in set C, and select the minimum value d among these distances.

Step 4: Compare T_1 with distance d. If d < T_1, add the sample point Q to the canopy with the smallest distance from it, and label a weak marker. If d > T_1, set Q as the new canopy distance center, add sample point Q to set C, and remove Q from dataset X.

Step 5: Compare T_2 with distance d. If d<T_2, attach a strong label to it, update the center position of all strongly labeled samples to the cluster center of this canopy, and delete the sample point Q from the dataset X.

Step 6: Repeat steps 3, 4 and 5 until dataset X is empty.

Output: Cluster center set C

4. EXPERIMENT AND ANALYSIS

4.1 Data collection

This paper desensitizes the collected student behavior data. The original data includes the one-card consumption data of some undergraduates in a university in Henan, library access control data and educational administration system data. For details, see Table 1-3.

Table 1. Student card canteen consumption data.

Student ID	Name	Amount	Overage	Settlement department	Time
*******0135	Zhang**	9.0	161.5	Chicken soup	2019/3/12 11:24
*******0145	Wu*	5.5	116.0	Meat pie	2019/3/12 12:06
*******0124	Li*	10.5	65.0	Duck leg rice	2019/3/12 11:45

Table 2. Library loan data.

Date	Operation type	Book title	Index number	Reader ID
2019/3/23 16:21	Lend	Principles of Education	G40/323	1540*******
2019/3/24 8:54	Lend	Counseling Psychology	C932/42	1540*******
2019/3/24 9:15	Lend	Chrysanthemum and the Sword	H3189/3927	1540*******

Table 3. Academic affairs system grade data.

School year	Semester	Student ID	Name	Course title	Score	GPA	Retest mark
20192020	1	********0136	Peng*	Machine learning	86	3.6	
20192020	1	********0137	Wang**	Machine learning	69	2.9	
20192020	1	********0138	Li**	Machine learning	55	0	80

4.2 Data processing

The raw data of students' behavior recorded in various information systems of the school are too large and cumbersome to directly mine knowledge. In this paper, statistical methods are used to compress, generalize and normalize the student behavior data to make the data more valuable for analysis. For example, for the student card data, this paper first uses statistical methods to calculate the average monthly consumption amount, consumption frequency and consumption peak of students using a monthly cycle. After preprocessing, the evaluation indicators of students are shown in Table 4.

Table 4. Student evaluation metrics.

Dimension	Indicator name	Describe
Consumption law	Amount of consumption	Average monthly spending by students
	Consumption frequency	Average number of purchases by students per month
	Single consumption	Average single spending by students
	Peak consumption	Maximum student monthly spending
Living habits	Online time	Average monthly time spent online
	Work and rest rules	Average number of early wakes per month
Learning situation	Number of library visits	Average number of visits to the library per month
	Number of books borrowed	Number of books borrowed from the library
	Average scores	Grade point average per semester
	Course pass rate	Number of courses passed/Number of all courses
	Class attendance	Attendance times/attendance times

4.3 Experimental result

Due to space limitations, here we only use canopy and k-means algorithm to cluster student consumption data, analyze students' consumption patterns, and construct the consumption portrait of students on campus. The result is shown in Table 5 .

Table 5. Clustering results of students' consumption.

Number	Average monthly consumption	Monthly peak consumption	Monthly consumption frequency	Average monthly single consumption	Percentage of students
1	237.5	317.0	37.8	6.3	24.2%
2	523.6	562.2	113.2	4.6	40.9%
3	840.7	910.5	121.4	6.9	16.2%
4	391.4	423.9	96.3	4.1	18.7%

The monthly average consumption of students in group 1 is the lowest among all groups, but it cannot be simply judged that the overall consumption level of this group of students is low. The consumption frequency of such students is lower, and the average single consumption is higher among all groups. Therefore, it can be inferred that such students like off-campus consumption, and the overall consumption level is relatively high.

For the second group of students, the peak monthly consumption of these students is relatively close to the average monthly consumption, and the average monthly consumption and single consumption are in the middle level of each group. According to the consumption characteristics of such students, it can be seen that their consumption is relatively regular, basically solved in schools, and the overall consumption level is at a moderate level.

Group 3 students make up 16% of the total. Average monthly consumption, monthly consumption frequency and single consumption are relatively high. It can be inferred that their overall consumption level is relatively high and their living expenses are relatively sufficient.

Group 4 students, the average monthly consumption and single consumption are lower. It can be seen that the overall consumption level of such students is relatively low, and they belong to economically poor students. Schools should prioritize these students in financial aid evaluations.

4.4 Student portrait

According to the clustering results of the three indicators of consumption law, life law and effort level, the student categories with different behavior characteristics are summarized, categorized, and labeled, and the label of each indicator and the basic information of the student are integrated to describe the student Behavioral portraits. The portrait of the student is shown in figure 4.

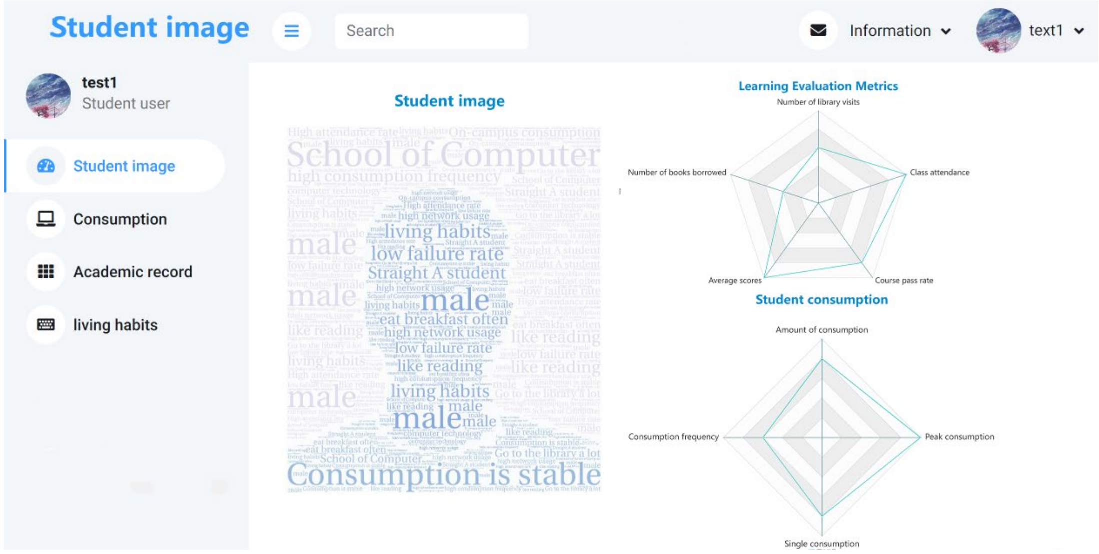

Figure 4. Student image.

5. CONCLUSION

The college student portrait system uses the relevant components of the Hadoop framework to integrate, clean and store the student data, and uses k-means and canopy algorithm to perform cluster analysis on the processed student data. The system constructs student portraits from the perspectives of student consumption patterns, life patterns, and effort levels, and objectively displays the characteristics of student groups. The follow-up work can analyze teachers, build teacher portraits, understand the current behavior status of teachers from teachers' personal information, scientific research results, and teaching evaluation, and provide quantitative decision-making basis for personnel management work.

REFERENCES

[1] Muhamad, W., Kurniawan, N. B., Suhardi and Yazid, S., “Smart campus features, technologies, and applications: A systematic literature review,” 2017 International Conference on Information Technology Systems and Innovation (ICITSI), 384-391 (2017).

[2] Valks, B., Arkesteijn, M. H., Koutamanis, A. and Heijer, A. C. D., “Towards a smart campus: Supporting campus decisions with Internet of Things applications,” Building Research & Information, 49(01), 1-20 (2021).

[3] Fischer, C., Pardos, Z. A. and Baker, R. S., “Mining big data in education: Affordances and challenges,” Review of Research in Education, 44(1), 130-160 (2020).

[4] Zheng, Y. F., “Survey of big data visualization in education,” Journal of Frontiers of Computer Science and Technology, 15(03), 403-422 (2021).

[5] Dwivedi, S. and Roshni, V. S. K., “Recommender system for big data in education,” 2017 5th National Conference on E-Learning & E-Learning Technologies, 1-4 (2017).

[6] Viloria, A., Naveda, A. S. and Palma, H. H., “Using big data to determine potential dropouts in higher education,” Journal of Physics: Conference Series, 1432(1), 012077 (2020).

[7] Khan, A. and Ghosh, S. K., “Student performance analysis and prediction in classroom learning: A review of educational data mining studies,” Educ. Inf. Technol., 26, 205–240 (2021).

[8] Yin, Y., Yu, X., "Portrait of college students and personalized teaching mode under the blended learning mode based on big data technology," International Conference on Cognitive based Information Processing and Applications (CIPA), 992-997 (2022).

[9] Grivokostopoulou, F., Perikos, I. and Hatzilygeroudis, I., "Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance," 2014 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), 488-494 (2014).

[10] Liu, K. S., Ni, Y. K., Li, Z. and Duan, B., "Data mining and feature analysis of college students" campus network behavior," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), 231-237 (2020).

[11] Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G. and Rangwala, H., "Predicting student performance using personalized analytics," In Computer, 61-69 (2016).

[12] Ge, S. H., Wan, Q. and Bai, C. J., "Hadoop-based college student behavior warning decision system," Computer Applications and Software, 38(01), 6-12 (2021).