BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

Research and Application on Spark Clustering Algorithm in Campus Big Data Analysis

Qing Hou* **Guangjian Wang** **Xiaozheng Wang** **Jiaxi Xu** **Yang Xin**

Nanjing Xiao Zhuang University, Jiangsu, Nanjing, 210017, China

ARTICLE INFO

Article history
Received: 14 January 2020
Accepted: 17 January 2020
Published Online: 31 March 2020

Keywords:
Spark
Clustering algorithm
Big data
Data analysis
Mllib

ABSTRACT

Big data analysis has penetrated into all fields of society and has brought about profound changes. However, there is relatively little research on big data supporting student management regarding college and university's big data. Taking the student card information as the research sample, using spark big data mining technology and K-Means clustering algorithm, taking scholarship evaluation as an example, the big data is analyzed. Data includes analysis of students' daily behavior from multiple dimensions, and it can prevent the unreasonable scholarship evaluation caused by unfair factors such as plagiarism, votes of teachers and students, etc. At the same time, students' absenteeism, physical health and psychological status in advance can be predicted, which makes student management work more active, accurate and effective.

Chinese Library Classification: TP311 Document code: A

1. Introduction

By 2013, big data had penetrated into all fields of society and brought about profound changes [1]. Big data is a great change in thinking, a source of power for human beings to acquire new cognition and create new values, and a method for changing the market and innovating educational management. It is not only a technology, but also a value and methodology. By mining, analyzing and synthesizing the data, more valuable products and services can be obtained. In China, there are many colleges and universities with more than 10,000 students and teachers. For university management, a large number of information data will be generated through the Internet login and meal card consumption, such as student's information, course selection, school report card, book borrowing history, online time distribution, internal forum communication, MicroBlog and WeChat, etc. The existing huge information system in these universities has accumulated a lot of basic original data through years of operation. Carrying out in-depth analysis and application of these original data, strengthening the scientific management of the school based on overall analysis, and offering data support for the development decision of the school, has become an important issue and pioneering opportunity for Chinese universities. At present, data analysis and application in colleges and universities are mainly used to assist teaching management in many aspects such as

**Corresponding Author:*
Qing Hou,
NanJing XiaoZhuang University, Jiangsu, Nanjing, 210017, China;
Email: 815422078@qq.com
Fund Project:
Nanjing Key Laboratory of Intelligent Information Processing Open Fund Project (No.19AIP05)

DOI: https://doi.org/10.30564/jcsr.v2i1.1808

scientific research calculation, enrollment promotion, subject management, overall salary planning, and student information tracking. Data mining in student management is rare . In the management of college students, managers can know students' static data (such as grades, curriculum, personal basic information, etc.), but it is difficult to master the dynamic data for students' behavior. If big data analysis on campus is related to student behavior analysis, such as consumption of three meals a day, class attendance, library access and reading, daily consumption (bathing, school supermarket, printing consumption, school hospital consumption), etc., then some students' learning conditions, physical health conditions and mental health conditions can be excavated and analyzed. This offers forward-looking reference data for the management of college students and improves the accuracy and efficiency of management.

Big data analysis needs to rely on a well-performed data analysis platform. Traditional high-performance single-machine running big data is no longer practical. Several big data system analysis platforms, such as Hadoop, PureData and Exadata, have been launched in recent years, with Hadoop platform being the most prominent and popular among users. However, with the deepening of application, Hadoop has exposed its limitations. This is mainly reflected in the following aspects: first, the operation is too single and only supports Map and Reduce operations. Second, iterative computation is inefficient, especially in machine learning and graphic computation. These issues are better addressed by the Spark framework technology proposed by the Apache Software Foundation at the end of 2013. Spark is a parallel computing architecture based on HDFS. The main idea is to reduce disk and network I/O overhead through a new job and data fault-tolerant approach whose core technology is elastic distributed data sets (RDD). Unlike MapReduce, Spark is not only limited to writing map and reduce. It offers users with a more powerful memory computing model, enabling users to read data into the cluster's memory through programming, which can quickly iterate data sets in memory for many times and support complex data mining algorithms and graph computing algorithms. At present, Spark has built its own whole big data processing ecosystem, such as flow processing, graph technology, machine learning, NOSQL query, etc., and is a top-level Apache project. Although Spark requires high memory and its launch time is relatively short , with the gradual maturity of big data related technologies and industries, Spark technology has developed rapidly with unparalleled advantages after Hadoop and will become the next generation of cloud computing and big data core technology to replace Hadoop [2]. Therefore, it is a good choice to analyze campus big data with Spark related technologies.

This paper, mainly based on Spark distributed platform, analyzes the daily behavior of university students by collecting their smart card data with clustering algorithm and from multiple angles. To provide more accurate and effective data for the evaluation of scholarship may benefit those who actively work hard and restrain the unfairness of plagiarism and fraud votes.

2. Valid Data Sources

The data sources used in this research are all-in-one card data of my University in 2014. It mainly include: canteen card swiping information (including breakfast, lunch and dinner), supermarket card swiping information, learning and consumption card swiping information (printing and copying, book purchase, etc.), bathhouse card swiping information, daily life consumption card swiping information (such as haircut, etc.), totaling more than 20,000. It is essential that these data be filtered out to establish a relationship with the target issue. Scholarships are generally assessed in two ways. One is an ordinary scholarship with good results, and the other is an inspirational scholarship with good results and poor families. For ordinary scholarships, it is unfair to simply look at the results or the election of teachers and students. For example, students may cheat and lead to better results. They may also get the relevant examination questions, or they may not work hard at ordinary times and finally suddenly recite them. In addition to the conditions for scholarships, inspirational scholarships also have a requirement for family difficulties. Student managers can only understand family difficulties through their own descriptions and the relevant unit certificates offered by the students, but do not know that such certificates are often easy to fake or find a relationship seal to get. Therefore, if students' daily behaviors are analyzed through all-in-one card data information and the scope of personnel who actively work hard in learning is locked, then the student management personnel can choose from this scope according to their learning achievements, which will achieve the relatively fair goal.

Generally speaking, scholarship recipients should start and finish classes on time, study hard and have good results. Students' daily behavior is that students eat breakfast and lunch by the hour, which accounts for a large proportion of learning and spending, and they often go to and from the library. The card information shows the time of swiping the card when students go for breakfast, lunch and dinner. In this work, the average number of breakfasts and lunches in a month except weekends

 DOI: https://doi.org/10.30564/jcsr.v2i1.1808

is counted to indicate whether students attend classes on time, in which breakfast time is between 6:00 and 7:40, and lunch time is 11:40 to 12:40. Therefore, the behavior analysis of motivational scholarship personnel is set from four dimensions: monthly average consumption, breakfast by point, noon by point and study consumption ratio. After screening, searching, removing and deleting invalid data, a total of 18,389 valid data are obtained, including: student number, total consumption, canteen consumption, study consumption, breakfast (6:00-7:40), lunch (11:40-12:40). The authors group students into different catalogs through multidimensional clustering: students who eat breakfast and lunch on time (they are not late or absent from school), students who often go to the library, and students who spend a lot of money on study but the overall consumption level is not too high. Such students can offer reference basis for schools to evaluate inspirational scholarships. The evaluation standard of ordinary scholarships do not need to care about canteen consumption

3. K-means Clustering Algorithm

3.1 Algorithmic Thinking

The basic idea of K-means algorithm is to give K cluster centers randomly in the initial state and divide the sample points to be classified into clusters according to the nearest neighbor principle. Then, the centroid of each cluster is recalculated according to the average method to determine the new cluster center. Iterative until the moving distance of cluster center is less than a given value. K-means clustering algorithm is divided into the following three steps [3]:

(1) The first step is to find a clustering center for the points to be clustered.

(2) The second step is to calculate the distance from each point to the cluster center and cluster each point into the cluster nearest to the point.

(3) The third step is to calculate the coordinate average of all points in each cluster and use the average as a new cluster center. Step 2 and step 3 should be repeated until the cluster center no longer moves in a large range or the number of clusters meets the requirements.

Spark MLIib K-means clustering model uses K-means algorithm to calculate clustering center. Mllib implements K-means clustering algorithm: firstly, cluster centers are randomly generated, which supports randomly selecting sample points as initial center points, and also supports K-means++ method to select the optimal cluster center points. Then the center point of the sample is calculated iteratively. The distributed implementation of the iterative calculation focus is: firstly, each sample's center should be calculated; secondly, it should sum up the sample values and count the number of samples through aggregate function; finally, it should obtain the newest center point, and judge whether the center has changed.

3.2 Algorithm Implementation

K-means algorithm is one of the main algorithms in Spark Mllib algorithm. Spark-based distributed K-means algorithm offers a good practical tool for big data clustering [4]. The source code implementation of K-means algorithm is not repeated here. the following are the key codes for clustering analysis of campus data based on Spark Mllib K-means algorithm:

```
object Test1 {
  def main(args :Array[String]): Unit ={
    val conf = new SparkConf().setMaster("local").setAppName("Consume")
    val sc = new SparkContext(conf)
    Logger.getRootLogger.setLevel(Level.WARN)
    //val data = sc.textFile("/home/spark/file1/FLWING1_ medical.csv")
    val data = sc.textFile("/input/NJXZC_FLWing.csv")
    val data1 = data.map(d => d.split(","))
    val stu_no = data1.map(arr => arr(0))
    val consume_data = data1.map(arr => arr.slice(1,6))
     val parseData = consume_data.map(consume => Vectors.dense(consume.map(str => str.toDouble))).cache()
    val initMode = "kmeans//"
    val numClusters = 10
    val model = new KMeans()
      .setInitializationMode(initMode)
      .setK(numClusters)
      .setMaxIterations(numClusters)
      .run(parseData)
    for(c < -model.clusterCenters)
      println(c.toString)
    val a = parseData.collect
    val b = model.predict(parseData).collect
     val c = model.predict(parseData).map((_,2)).countByKey()
    c.foreach(println)
    val wssse = model.computeCost(parseData)
    println("sse =" + wssse)
  }
}
```

4. Experiments and Data Analysis

4.1 Experimental Environment

The experimental environment is Spark distributed cluster environment. One Dell server T410, hard disk 1.4T

DOI: https://doi.org/10.30564/jcsr.v2i1.1808

memory 100G, and 10 virtual machines are invented via Vmware. Each virtual machine has 8G of memory, 100G of hard disk, one processor and 32-bit ubuntu operating system. After Spark related configuration files and program installation, a Spark cluster with ten nodes was established.

There is 18,389 valid experimental data after screening, querying and removing invalid data from over 20,000 smart card related data. The data fields are: the average level for the four months from March to June, including student number, total consumption amount, canteen consumption amount, study consumption data, breakfast (6: 00-7: 40) and lunch time (11: 40-12: 40).

Average monthly total consumption = total consumption of 4 months card / 4

Proportion of canteen consumption = average monthly consumption / average monthly total consumption of canteens

Proportion of study consumption = average monthly study consumption / average monthly total consumption

Number of breakfast meals by times = Count the total number of breakfast meals in the range of 6:00-7:40 in 4 months except weekends / 4

Number of lunches by order = count the total number of breakfasts in the range of 11:40-12:40 in 4 months except weekends / 4

Figure 1 is a running effect of the clustering algorithm under Spark cluster. The running result shows the clustering number and the center of mass after clustering. Student number of each cluster can be output through routine call.

Figure 1. Spark cluster experiment output results

Table 1 shows the clustering results after multiple clustering optimizations.

4.2 Result Analysis

The student manager often sees static data, such as student's report card, student's election results. It is impossible to observe the student's behavior in all aspects, to monitor whether the student eats on time and attends class on time in real time, to print some review materials frequently, to prove extravagance and waste but with poverty certificate and to obtain inspirational scholarships. Through Table 1, whether there is a certain relationship between students who eat on time and study consumption is analyzed. With the increase of study consumption, the number of regular meals for breakfast and lunch also increased. This indicates that students who eat on time and attend classes on time and love study are usually active learners, as shown in Figure 2. After comparative analy-

Table 1. Clustering results after multiple clustering optimizations

Cluster number	Number of clusters	Average monthly total consumption	Average monthly canteen	Proportion of canteen consumption	Average monthly consumption of study	Ratio of study to consumption	Breakfast on time	Lunch on time
4	829	28.656	20.845	72.745%	2.367	8.259%	0.146	0.815
5	371	110.361	87.971	79.712%	3.981	3.607%	0.864	3.717
10	1572	192.278	156.559	81.423%	6.287	3.270%	2.049	6.863
0	2150	266.478	217.100	81.470%	8.885	3.334%	3.763	9.446
7	2112	330.180	269.131	81.510%	10.121	3.065%	4.809	11.316
8	1349	442.717	352.098	79.531%	12.025	2.716%	5.745	13.621
2	2106	380.554	321.805	84.562%	10.927	2.871%	5.802	13.049
3	102	479.555	413.211	86.165%	12.753	2.659%	6.058	14.917
11	2643	552.128	459.978	83.310%	16.262	2.945%	6.265	15.759
1	2133	911.025	776.607	85.245%	41.331	4.537%	6.450	22.350
9	1127	738.482	632.256	85.616%	22.149	2.999%	6.628	18.086
6	1894	623.159	536.504	86.094%	17.369	2.787%	6.734	17.096

 DOI: https://doi.org/10.30564/jcsr.v2i1.1808

sis, all students with cluster number 1 spend a lot of money on study and eat breakfast and lunch on time, which is the best locking range for scholarship evaluation. At the same time, the number of student canteens has a higher proportion of consumption and higher learning consumption, which is also the best locking range for motivational scholarship pacifiers. Using the campus card data, we can also explore whether students have friends and are lonely, and pay attention to students' psychological health. Students' physical health status is mined by swiping the campus hospital data with a all-in-one card. The brunch card time is used to predict the number of students who were frequently late and absent from school. As increasingly data sources (such as educational administration data, library data, WeChat and MicroBlog forum, etc.) are opened to the research group, our research content and accuracy will be improved.

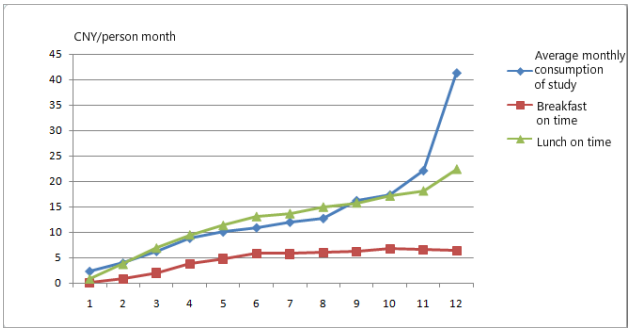

Figure 2. Partial consumption relationship diagram

5. Conclusion

In this paper, the information of all-in-one card for school students is taken as the research sample. Big data mining technology based on Spark platform is applied, combined with KMeans clustering algorithm. This paper takes scholarship evaluation as an example to analyze big data. Data includes analysis of students' daily behavior from multiple dimensions, and it can prevent the unreasonable scholarship evaluation caused by unfair factors such as plagiarism and fraud votes and can improve student management work. Further study can be extended to analyze useful data and be better-prepared for things like truancy, physical health and psychological status of students. All the efforts will provide more systematic and valuable management data for student management work, improve management efficiency and try to avoid the preventable in advance.

References

[1] Yihua Huang. Understanding Big Data[M]. China Machine Press, 2014.

[2] Meiling Huang. Spark MLlib Machine Learning: Algorithm, Source Code and Actual Combat Details[M]. Publishing House of Electronics Industry, 2016. (in Chines)

[3] Aiwu Zhou, Dandan Cui, Yong Pan. An Optimization Initial Clustering Center of K-means Clustering Algorithm[J]. Microcomputer and Its Applications, 2011, 30(13): 1-3.

[4] Weizhong Zhao, Huifang Ma, Yanxiang Fu, et al. Research on Parallel K-means Algorithm Design Based on Hadoop Platform[J]. Computer Science, 2011(10): 166-168.

[5] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[6] Jianpei Zhang, Yue Yang, Jing Yang, et al. Algorithm for Initialization of K-Means Clustering Center Based on Optimized-Division[J]. Journal of System Simulation, 2009, 21(9): 2586-2589.

[7] The Apache Software Foundation. Apache Mahout: Scalable Machine Learning and Data Mining [EB/OL], 2014.

[8] F Wang, Z Liu. Optimization method of distributed K-means algorithm based on Spark. Computer Engineering and Design, 2019; 40(6): 1595-1600. DOI: 10.16208/j.issn1000-7024.2019.06.017

[9] Y Qu, W Deng, F Hu, et al. Algorithm for ordering points to identify clustering structure based on spark. Computer Science, 2018; 45(1): 97-102+107. DOI: 10.11896/j.issn.1002-137X.2018.01.015

[10] M Xu, C Yu, H Shen. Research on K-means algorithm of spark parallelization. Microelectronics & Computer, 2018, 35(5): 95-99.

[11] Liu P, Teng J, Zhang G, et al. Parallel K-means algorithm for massive texts on spark. The 2nd CCF Big Data Conference, 2014. (in Chinese). Available from: http://mahout.apache.org/

DOI: https://doi.org/10.30564/jcsr.v2i1.1808