The Design of College Student Growth Management System Based on Data Mining

Jiaqi Yuan[1], Shi Cheng[1], Jian Lu[2]

[1] School of Computer Science and Technology, Nantong University, Nantong, Jiangsu, China, 226019

[2]Jiangsu Vocational College of Business, Nantong, Jiangsu, China, 226019

Keywords: management system based on data mining; college student growth; design

Abstract: The concept of “big data” has been mentioned much more times than ever before. At present, the hardware and distributed system with high performance can help us process the data more quickly, and handle “all data” instead of “random data”. It is not necessary to depend on the sampling method when dealing with all the data and the result will be more reasonable. College student growth management system is designed to solve the problems such as “what kind of job will suit me” or “what courses will bring troubles for me”. Employment guidance works by finding similar students in database by using collaborative filtering algorithm. Academic warning tries to find rules in the courses that students often fail with the classic Apriori algorithm.

1. Introduction

The number of college students increases every year and it has broken through 16.13million in China. By early 2017, there are thirty-eight comprehensive universities with more than 40,000 students. A sufficient number of students will provide us the data for analysis and prediction. The most important of all, every student’s record is real and valid. For example, there is realy very little chance that a student can cheat in an exam, so his transcript can represent his learning state. At present, colleges and universities use database to record students’ academic performance, by connecting to the data base, we can get a lot of information such as which subject the student is good at and which is bad. We can create various kinds of models to describe the learning state, by comparing to the models, we can provide appropriate learning plans for the student to improve himself. Employment guidance can also be provided by recording the employment of past graduates. Besides the education database, the Packaged Campus Card is essential to the growth management system. Benefit from the construction of digital campus, the campus cards are commonly used among university students. When the students come to the dining hall or the supermarket in school, they usually use the campus cards. So we can get the consumption data in the Campus Card data canter. By summarizing the consumption, we can collate the economic situation of the students, and decide if it is necessary to give the scholarship.

2. The Architecture of the System

The system is based on Hadoop[1] which is an open-source data processing framework that includes distributed data processing model and execution environment, named MapReduce[2], and distributed File System, named Hadoop distributed File System (HDFS). It is well received because of its economy and efficiency. The largest Hadoop-based cluster is installed at Facebook to manage nearly 31 PB of online disk data[3].The system can be deployed on several ordinary computers, files and the data can be shared by using the HDFS. In order to analyze the students’ data, we have to connect to Employment database, Education Databse and Campus Card Database. The architecture is shown in Figure 1.

In order to collect the structured and unstructured data from the existing traditional database, the Apache Sqoop technology will be used. The data will be stored relies on Hadoop HDFS and the Hive data warehouse.

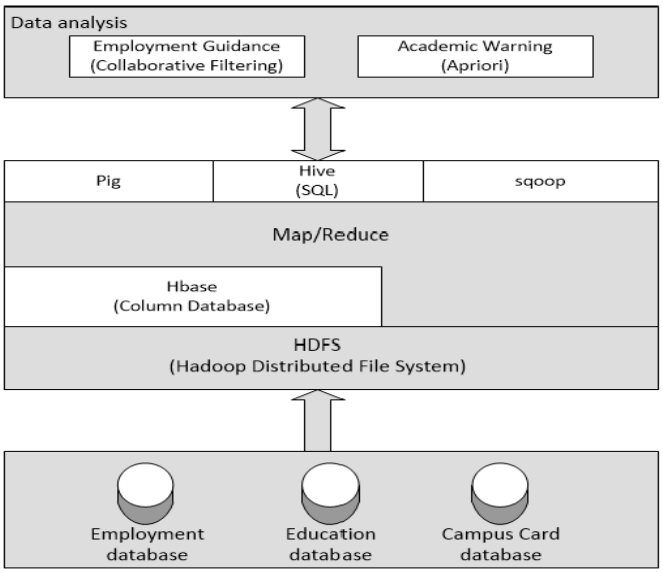

Figure 1. Architecture of the system

3. Main Functions Provided by the System

3.1 Major selection and employment guidance

When a senior high school student decide to study further in a college, It is very important to think carefully before selecting the major. Professional choice is bad, be equal to is misstep circle. In the circumstance a summary of the past graduates will be very helpful. The reports will show what the graduates do after graduation and they will be sorted by specialites (shown in Figure2).Enterprises are divided into two categories: excellent enterprises and normal enterprises. This will make great sense when a student is indecisive in choosing the major. The students will not feel confused about the future when they see the real data about their specialties.

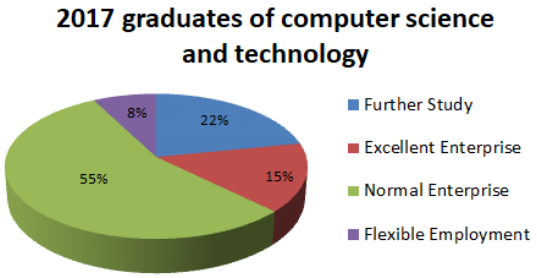

Figure 2. The situation about the graduates

Employment database is created to record the situation after the graduation of the students. Besides employment situation the database will also contain academic record, awards record, career record and main achievements. This information can be used to find the similar students. For example, Student i(has graduated several years) got a contract in a enterprise and have a good salary. Student j (will graduate in a few months) is similar to Student i, then we can provide j with the situation. If the enterprise appreciates what i has done and want to recruit more people like this, Student j will be recommended. Collaborative filtering algorithm is used to find the similar students. Collaborative filtering recommendation is the most successful strategy including User-based Collaborative Filtering(Memory-Based Collaborative Filtering),Item-Based Collaborative Filtering and Model-Based Collaborative Filtering[4].In order to find the similar students, User-based Collaborative Filtering is an appropriate algorithm. The similarity between student i and student j is

recorded as sim(i,j).The student's performance in each aspect can be seen as a M dimensional vector as Table1.C_1 to C_{M-3} represent compulsory courses,C_{M-2} represents avarege of optional courses, C_{M-1} represents awards and C_M represents the situation that the student take part in school activities.Range of value for each count is 0 to 10.

Table 1: Evaluation vector table for students

	C_1	C_2	… ….	C_{M-3}	C_{M-2}	C_{M-1}	C_M
$student_1$	8	7	… …	8	8	5	3
.....	… …	… …	… …	… …	… …	… …	… …
$student_i$	7	6		7	7	7	6
……	… …	… …	… …	… …	… …	… …	… …
$student_j$	7	6		7	8	6	5
……	… …	… …	… …	… …	… …	… …	… …
$student_n$	6	7	… …	8	8	4	4

The Adjusted Cosine Similarity formula is used to calculate the similarity between i and j.In the formula $I_{i,j}$ represents the set which record the scores about student i and j. $R_{i,c}$ represents the score student i got in project c, $\bar{R}_i$ represents the average score of student i and $\bar{R}_j$ represents student j.

$$\text{sim(i, j)} = \frac{\sum_{c \in I_{i,j}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_j} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_j} \left(R_{j,c} - \bar{R}_j\right)^2}}$$

In this way. the similarity between two students can be quantified, so we can give some constructive advice based on past data.

3.2 Academic warning for students

Nowadays, some college students do not work so hard in school so they have a lot of courses that can not be passed. They will lost their degrees if they can not get the credits for these courses or worst of all they will just get certificates after many years of learning. Academic warning Function is designed to avoid situation like this. The failed courses for every student will be recorded and we will try to find the relations among them. For example, the data shows that 82% of the students who failed to pass the course "Advanced Mathematics" have failed in "Advanced Programming Language" too. If a new coming girl has not pass the test of "Advanced Mathematics", we should advise her to avoid "Advanced Programming Language" if it is not compulsory. Or she should pay more attention to this course if it is required for the degree. The warning based on real data will attact students' attention. Besides the college courses, the relation between school curriculum and social examination can also be found by data mining.

We decide to do the research with Apriori algorithm[5].Let I=$\{i_1, i_2, \ldots \ldots i_n\}$ be a set of items. Let D=$\{t_1, t_2, t_3, \ldots \ldots t_m\}$ be a set of transactions constituting the database.Each transaction in D has a unique ID and contains a subset of the items in I.The form $X \Longrightarrow Y$ means a rule of association when $X, Y \subseteq I$ and $X \cap Y = \Phi$.Support($X \Longrightarrow Y$) means the proportion of transactions in D which contains the itemset($X \cup Y$),Confidence($X \Longrightarrow Y$) means the percentage of transactions including both X and Y in transactions which contains X.Support($X \Longrightarrow Y$)= $P(X \cup Y)$; Confidence($X \Longrightarrow Y$)=$P(Y|X)$ while P means probability.Apriori is an algorithm for mining frequent itemsets.When the minimum thresholds on Support and Confidence are given,we can find frequent itemsets in the

set of transactions.Every frequent itemset will be deal with to find the potential relationship.

In the academic warning system, Let itemset I be the set of college courses and D be the set of transactions which record the failed courses of every student. Let $i_1$to i_5 represent courses{“Advanced Mathematics”,”Data Structures”,”Object-Oriented Programming”,”Computer Network”,”Assembly Language”}.Table2 below represents transactions set D.The minimum Support is setted to be 50% and and minimum Confidence is 70%.

Table 2 Transactions Set D

Tid	Itemset
1	i_1,i_2, i_3, i_5
2	i_1,i_2, i_3,
3	i_3, i_5
4	i_1,i_2, i_4,

The algorithm will be described as follows:

(1) Candiate-set C1=$\{i_1, i_2, i_3, i_4, i_5\}$

(2) Calculate the Surpport of every item in C1 by Scaning the set of transactons D.The result will be 75%,75%,75%,25%,50%.Prune i_4 because its Surpport is less than the threshold.Frequent itemset L1=$\{i_1, i_2, i_3, i_5\}$

(3) C2 is generated by L1,C2=$\{\{i_1, i_2\}, \{i_1, i_3\}, \{i_1, i_5\}, \{i_2, i_3\}, \{i_2, i_5\}, \{i_3, i_5\}\}$.

(4) Calculate the Surpport of every item set in C2,the result will be 75%,50%,25%,50%,25%,50%.Prune the items in the same way,the new frequent item-set L2=$\{\{i_1, i_2\}, \{i_1, i_3\}, \{i_2, i_3\}, \{i_3, i_5\}\}$.

(5) C3 is generated by L2,C3=$\{\{i_1, i_2, i_3\}, \{i_1, i_2, i_5\}, \{i_1, i_3, i_5\}, \{i_2, i_3, i_5\}\}$,in C3,the item$\{i_1, i_2, i_5\}$ will be dropped because its sub set$\{i_2, i_5\}$ does not exist in L2.The items $\{i_1, i_3, i_5\}$,$\{i_2, i_3, i_5\}$ will be deleted because of the same reason. Finally C3=$\{i_1, i_2, i_3\}$.

(6) Calculate the only item in C3,the Support is 50% which is not less than the threshold,so L3=$\{i_1, i_2, i_3\}$.

(7) L= L1 ∪ L2 ∪ L3 = $\{\{i_1\}, \{i_2\}, \{i_3\}, \{i_5\}, \{i_1, i_2\}, \{i_1, i_3\}, \{i_2, i_3\}, \{i_3, i_5\}, \{i_1, i_2, i_3\}\}$,this is the frequent itemset or large itemset.

(8) The itemsets with a length more than 2 such as $\{i_1, i_2, i_3\}$ will be calculated. Confindence of all its subsets will be calculated. If the value is above 70%,a new association rule is found.

Confidence($\{i_1, i_3\} \Longrightarrow i_2$)=100% means that there is a great probability of failing to pass ”Data Structures” if a student failed in both “Advanced Mathematics” and ”Object-Oriented Programming”.

4. Conclusion

College student growth management system is designed to help students grow up. Employm-entguidance and academic warning are two main functions of the system. Calculations of massive amounts of data can help us to give some useful suggestions withwhich a student can be better motivated to lead a meaning life so that they can contribute to the society[6].Instead of considering only academic parameters, we could also consider personal character-istics, behavior, family history, past records in order to predict the performance of a studet. Some information can be acquired from the Campus Card data base, but how to create themathematical model is still need to undertake further analysis and research.2017 Jiangsu Province modern educational technology research topic (serial number: 2017-R-54131)

Acknowledgement

Higher education research in Nantong University (serial number: 2015GJ006)

References

[1] Apache hadoop 2009 http://hadoop.apache.org/

[2] Jeffrey D and Sanjay G 2008 Mapreduce: simplified data processing on large clusters Commun. ACM vol 51 pp 107-113

[3] Thusoo A et al. 2010 Data warehousing and analytics infrastructure at facebook ACM SIGMOD conf pp 1013-1020

[4] Wang guoxia, LIU heping. Survey of personalized recommendation system. Computer Engineering and Applications,2012,48(7):66-76

[5] Parack, Suhem, Zain Zahid, and Fatima Merchant 2012 Application of data mining in educational databases for predicting academic trends and patterns IEEE International Conference on Technology Enhanced Education (ICTEE)

[6] Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer, William F. Punch 2003 Predicting student performance: an application of data mining methods with an educational web-based system 33rd ASEE/IEEE Frontiers in Education Conference