Research on Student Behavior Inference Method Based on FP-growth Algorithm

Jingwen Li [1,2], Na Yu [2], Jianwu Jiang [1,2,*] , Xu Li [2], Yuan Ma [2], Wenda Chen [2]

[1] Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin University of Technology, Guilin 541004, China
[2] Guilin University of Technology, Guilin 541004, China

KEY WORDS:FP-growth；Association rule；Frequent itemsets；Intelligent Recommendation；Collaborative filtering algorithm

ABSTRACT:

How to use modern information technology to efficiently and quickly obtain the personalized recommendation information required by students, and to provide high-quality intelligent services for schools, parents and students has become one of the hot issues in college research. This paper uses FP-growth association rule mining algorithm to infer student behavior and then use the collaborative filtering recommendation method to push information according to the inference result, and then push real-time and effective personalized information for students. The experimental results show that an improved FP-growth algorithm is proposed based on the classification of students. The algorithm combines the student behavior inference method of FP-growth algorithm with the collaborative filtering hybrid recommendation method, which not only solves the FP-tree tree branch. Excessive and collaborative filtering recommendation algorithm data sparseness problem, can also analyze different students' behaviors and activities, and accurately push real-time, accurate and effective personalized information for students, to promote smart campus and information intelligence The development provides better service.

1. INTRODUCTION

In recent years, with the development of "digital campus" and "smart campus" in colleges and universities, the core role of card in campus has become more and more prominent [1]. As a must-have item for teachers and students, the campus card records related data such as student dining, borrowing books, supermarket access control, and online data. These data hide the information of most students' daily behaviors, and through data mining and Analysis of students' daily behavior can inform students' academic status in advance [2]. In the current research field, foreign research scholar Kunyanuth [3] analyzed the students' learning behavior through data mining technology, and constructed a student behavior analysis model system to help students develop more effective and more appropriate learning methods. Zhao et al. [4] constructed a student learning model by analyzing the data of students' online learning, and provided personalized help to students while assessing and classifying students' learning ability. Domestic research scholars Jiang Nan and Xu Weisheng[5] analyzed the students' learning and consumption behaviors through data on consumption and access control through improved K-means and Apriori algorithms, and provided better students by mining the related information between students' learning and behavior. Service. Ji Zhen [6] and others analyzed the consumption data of each student in the campus card to provide student information for the school catering operators, which further improved the management and service level of the school catering industry. At present, the correlation algorithms of data mining and data correlation can analyze the associated data, but these algorithms have corresponding defects. Therefore, this paper proposes a behavior inference method based on FP-growth algorithm for the defects of the above algorithms. Research method, which analyzes popular activities/information and high-quality activities/information through log analysis, records relevant data information, and uses the improved FP-growth association rule mining algorithm to push information inference results to users in real time. The algorithm is validated by the student behavior example of the school. Based on the student classification, the method combines the student behavior inference method based on FP-growth algorithm with the collaborative filtering hybrid recommendation method, which solves the problem of excessive FP-Tree branching. Collaboratively filtering the recommendation data sparseness problem while providing students with real-time, effective personalized activities/information push that meets their needs.

2. LOG ANALYSIS

Logging is a very useful source of information for computer system resource management users and application management and security [7]. Log analysis is to record and summarize related information or behavior browsing and summary times of users clicking on the active platform. Log analysis can be used to derive activities and information that are of interest to the user, facilitating subsequent recommendations. Log analysis can be used as a source of activity information for smart push services, and can obtain information such as user classification and user attributes. Therefore, by utilizing the complicated data information in the application, by logging the log, according to the activity situation, when the user uses the related application for publishing the event/information, the application records the user's data when browsing the activity/information on the application. The information thus pushes the corresponding activity/information for the user and provides data support for user behavior push mining.

2.1 Top Events / Information Analysis

The popular activity/information is to analyze all the activities and information in the application, and analyze the attributes such as the number of searchers, the number of visitors, the number of participants, etc. to obtain activities/information with more user needs. When the smart push service is provided to the user, the current hot

* Jianwu Jiang - E-mail: fengbuxi@glut.edu.cn

event/information can be prioritized, which improves the timeliness of smart push.

For the judgment of popular events/information, choose to evaluate according to the number of searchers, number of visitors, number of participants, etc. The evaluation rules are as shown in the table:

Activity evaluation	Weights（ $w_1 + w_2 + w_3 + \cdots w_n = 1$ ）
Number of searchers P	w_1
Number of visitors S	w_2
The number of participants J	w_3
…	…
Attributes N	w_n

Table 1 Top activity evaluation

The number of people searching and browsing determines the current popularity of the event/information, and the number of participants determines the popularity of the event, and a summary of these can be used to get the total heat of the event/information.Set the weight to .Therefore, we can evaluate its popularity X:

$$X = P * w_1 + S * w_2 + J * w_3 + \cdots + N * w_n$$

By calculating the popularity X of each activity/information, the higher the X, the more popular the activity/information is, and the more it should be pushed.

2.2. Quality activities / information analysis

Similar to popular events/information is quality events/information. For high-quality events/information, its richness, number of participants and subsequent acclaim are all high-quality, so according to the richness of event information, activities The number of participants, the praise of the event, etc. are analyzed and evaluated. The evaluation rules are as shown in the table:

Activity evaluation	Weights（ $w_1 + w_2 + w_3 + \cdots w_n = 1$ ）
Activity richness S	w_1
Number of participants J	w_2
Activity rating P	w_3
…	…
Attributes N	w_n

Table 2 Evaluation criteria for quality activities

The attribute values in the table are quantized to facilitate calculation, wherein the activity richness S and the activity popularity P can be expressed by Boolean values. Activity richness is determined by the amount of activity content, and each item has a count plus one to get the richness of the activity. The popularity of the event is determined by the praise given by the user. Each time there is a favorable record, the count is incremented by one, so that the popularity of the event is obtained. At the same time, these attributes are weighted as needed.The weight of the setting is , and thus, the quality Y of the activity/information is:

$$Y = S * w_1 + J * w_2 + P * w_3 + \cdots + N * w_n$$

By calculating the quality Y of each activity/information, it can be concluded that the higher the Y, the higher the activity/information among the students, the more it should be pushed; otherwise, the push is not.

3. FP-GROWTH ALGORITHM AND ITS IMPROVEMENT

3.1 User interest information acquisition and representation

The FP-growth algorithm is an association rule mining algorithm proposed by Han in 2000. The main advantage of this algorithm is that it does not generate candidate sets and only needs to traverse the data sets twice, which greatly improves the efficiency of mining frequent itemsets [8]. The FP-growth algorithm does not need to generate a candidate set, but uses a divide-and-conquer approach to directly generate frequent itemsets [9]. The Frequent Pattern Tree (FP-tree) is used to mine the relevance of frequent itemsets in the data. The data is judged once by the correlation in the data to determine whether there is a high correlation between the data. The high correlation of data, improve the accuracy of judgment between behaviors and the efficiency of mining frequent itemsets [10]. The FP-growth algorithm [11, 12] process is mainly divided into two aspects:

(1)Establishment of FP-Tree: First, the data set is scanned for the first time, and a frequent item set for counting the support in the data set is compiled. Then, according to the support degree of each item, a support threshold is set to perform item screening, and the purpose is to delete the item whose support degree is less than the threshold to increase the frequency of the item, and all the items of the items larger than the support threshold are called frequent 1- Item set. The items in the frequent 1-item set are sorted in descending order according to the support degree of the item to construct the item header table. Finally, the database is read twice, and the infrequent itemsets are deleted, that is, the items that are not in the frequent 1-item set, and the remaining items are inserted into the root node root in descending order to construct the FP-Tree.

(2) FP-Tree mining: FP-Tree mining starts from the least frequently used item of FP-Tree. Through the traversal of FP-Tree, you can find all the prefix paths in FP-Tree. It is the set of all items traversed by the set node to the root node, and the prefix path is the condition pattern base corresponding to the item set. Then, with the minimum support threshold value as the standard, the item node whose support degree is less than the threshold is deleted in the condition pattern base of the item, and the condition FP-Tree of the item is obtained. Finally, according to the condition FP-Tree, we can perform recursive mining on the full-array combination until the root node root is mined, and the frequent pattern is generated. If the conditional FP-Tree has only one path, the path can directly generate the frequent pattern.

3.2 Improved FP-growth algorithm

The traditional FP-Growth algorithm needs to recursively mine the frequent pattern base to mine the conditional FP-Tree, and the tree structure is stored by data structures such as linked lists. This storage method requires a large amount of memory,

especially for mining large-scale FP- In Tree, when FP-Tree has too many branches or the length of the branch is too long, constructing FP-Tree and traversing condition FP-Tree will consume a large amount of storage and computation memory, which greatly reduces memory utilization and thus greatly reduces The efficiency of mining data. Aiming at the defects of FP-growth algorithm in the research of the paper, this paper proposes an improved FP-growth algorithm. It mainly improves the behavior inference model of FP-growth algorithm for the problem of data sparsity of FP-growth algorithm. Because of the problem of data sparsity and data sparseness of special users/activities, the collaborative filtering push method will Collaborative filtering intelligent push method and FP-growth algorithm are combined to construct a new improved FP-growt algorithm recommendation model, which can not only collaborate to filter the problem of sparse data of intelligent push algorithm, but also solve the problem of excessive branch of FP-growth algorithm and occupy memory. Large and inferior data mining inefficiencies. Based on the behavioral inference results of the FP-growth algorithm, the collaborative filtering and hybrid push model is used to push the personalized information of the group, thereby further improving the effectiveness, accuracy and timeliness of the push results. The specific process is described as follows:

(1)Classification and aggregation according to the constraints in the information.

(2)Inferring according to the corresponding information entered in the classified database. Since the data information has been completely divided, the number relationship is more tight and the subsequent construction of the FP-Tree will not be too complicated.

(3)Setting a minimum support threshold and performing a scan on the required behavior database to generate a frequent item set to exclude behaviors whose behavior support is less than the minimum support degree, and sorting the eligible behaviors by number of times to generate a head set.

(4)Perform a secondary scan on the database according to the item set to construct the FP-Tree. If there is a common prefix, the prefix count is incremented by one. If there is no common prefix, the new node is assigned and the value is 1 and the FP-Tree is continuously constructed.

(5)After the FP-Tree is completed, the FP-Tree is mined to establish a behavior prediction table. Since the behaviors are in order, it is only necessary to mine the latter item to the FP-Tree based on the current behavior.

(6)Joint analysis of the resulting behavior mining table and database constraints to obtain the final behavior inference results.

The specific flow chart is as follows:

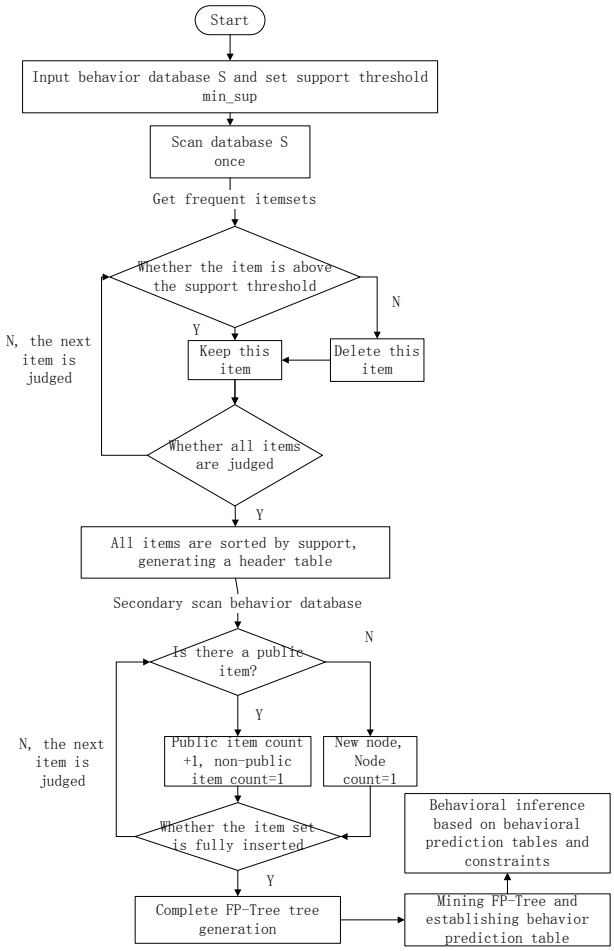

Figure 1 Improved FP-growth algorithm flow chart

4. EXPERIMENTAL ANALYSIS AND VERIFICATION

The Dream Space App is a student-specific app developed by the school. The app enriches students' after-school life and expands students' horizons by publishing the corresponding activities of colleges and universities. Therefore, the dream space platform can be used as a source of activity information for the smart push service, from which student-related information and resources are obtained. The experiment uses the dream space as the information source of the experimental data, and obtains a data set of the behaviors of the nine students based on the classification of the sports students from the platform, which has a library, classes, activities in the dormitory, and ongoing Activity data sets for club activities, sports, etc., as shown in the following table:

student	Behavior set
1	Library, class, dormitory
2	Class, club activities
3	Class, outdoor sports
4	Library, class, club activities
5	Library, outdoor sports
6	Class, outdoor sports
7	Library, outdoor sports
8	Library, class, outdoor sports, dormitory
9	Library, class, outdoor sports

Table 3 Sports Student Behavior Set

(1) The student behavior data data set is processed into a frequent item set to obtain a frequent

item set of student behavior. At the same time, set the minimum support threshold to 2, and the frequent item set is as shown in the table:

library	Class	Outdoor sports	social activity	dorm room
6	7	6	2	2

Table 4 Student behavior frequent itemsets

(2)Sort the frequent itemsets of student behaviors according to the number of frequent times, and delete the items below the minimum support threshold according to the set minimum support threshold 2, thus obtaining an orderly new frequent. The project set, which is the student behavior item head table, is shown in the table:

Class	library	Outdoor sports	social activity	dorm room
7	6	6	2	2

Table 5 support threshold is 2 student behavior item header table

(3)Returning to the student behavior data set, reordering the student behavior data set according to the item head table in (2), and obtaining a new student behavior data set compiled according to the item head set, and reordering the student behavior data. The set is as shown in the table:

student	Behavior set
1	Class, library, dormitory
2	Class, club activities
3	Class, outdoor sports
4	Class, library, club activities
5	Library, outdoor sports
6	Class, outdoor sports
7	Library, outdoor sports
8	Class, library, outdoor sports, dormitory
9	Class, library, outdoor sports

Table 6 Reordered Student Behavior Dataset

(4)Construct an FP-Tree tree based on the new student behavior data set. Root is used as the root node, and then the items in the student behavior data set are inserted one by one into the FP-Tree with root as the root node. If the previous item is used as the parent node, the child node is the latter item. If the inserted item has a shared node, all the shared nodes are incremented by one. If there is no shared node, the new node is inserted again. When all the data is inserted, FP-Tree is considered to be completed. The completed FP-Tree is shown below.

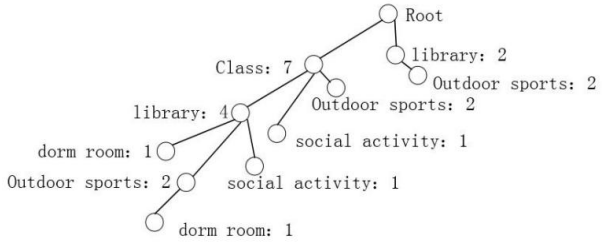

Figure 2 Student behavior inference FP-Tree

(5)After the FP-Tree is built, the FP-Tree is mined. Since the behavior is sequential, mining FP-Tree is not based on the least frequent items in the item header table, but is the closest to the root node, and the item with the highest count is backwards and summarized, and finally with minimum support. The threshold is a standard, and the result of the support less than the threshold is deleted, and the behavior prediction is completed. The behavior prediction table established based on the mining results is shown in the following table:

	Mining results	Behavior prediction
Class	{Class，library：4} {Class，Outdoor sports：2} {Class，social activity：1}	library，Outdoor sports
library	{library，Outdoor sports：4} {library，dorm room：1} {library，social activity：1}	social activity
Outdoor sports	{Outdoor sports，dorm room：1}	no

Table 7 Student Behavior Prediction Table

The above table is a classification of student movements, so the students who infer the behavioral class and go to the library will next carry out outdoor sports, while the outdoor sports have no behavior prediction and the frequent items are predicted to be outdoor sports. The behavior of setting outdoor sports is inferred as no. The behavior of the nine students on the second day is shown in the following table:

Serial number	Project set
1	Class, outdoor sports, dormitory, class
2	Library, dormitory, class, dormitory
3	Dormitory, class, outdoor sports, watching movies
4	Outdoor sports, dormitory, library
5	Class, outdoor sports, dormitory
6	Class, library, outdoor sports
7	Dormitory, library, outdoor sports
8	Library, class, outdoor sports, dormitory
9	Class, outdoor sports

Table 8 Student behavior data set for the second day

Among them, 63% of the situation after class and library behavior is outdoor sports.

(6) Recommend outdoor sports-related activities to them according to the activity attributes, such as: the current activities of the stadium, if the score is higher, push, the stadium's idle information, and so on. Finally, students are graded based on the activities they participate in to facilitate follow-up recommendations.

5. IN CONCLUSION

In the research process, the student behavior inference is taken as an example. After the log analysis of popular activities/information analysis and high-quality activities/information analysis, the FP-growth algorithm is integrated with the collaborative filtering intelligent recommendation algorithm, and a collaborative filtering based on student-activity is proposed. The push method avoids the data sparseness and new user/activity problems of the collaborative filtering method to a certain extent, improves the accuracy of the activity intelligent push, and on this basis, according to the activity push object hobby pushing service for evaluation, the most consistent Services that push object needs and preferences are automatically pushed to students. The method further improves the accuracy and effectiveness of the push results.

6. REFERENCES

[1] XIA Yang,WANG Fang.Construction and application of campus big data analysis platform based on campus card data[J].Journal of Central China Normal University(Natural Sciences),2017(S1):146-151.

[2] Guo Peng. Association of students' consumption behavior and performance data based on campus card [D]. Northwest A&F University, 2019.

[3] Kularbphettong K. Analysis of Students' Behavior Based on Educational Data Mining[J]. 2017.

[4] Zhao L X, Li R, Ye J M, et al. Analysis of Students' E-learning Behavior Based on Bik-Means Clustering Algorithm[C]. International Conference on Computer Networks and Communication Technology. 2017.

[5] Jiang Nan, Xu Weisheng.Student Consumption and Study Behavior Analysis Based on the Data of the Campus Card System[J]. Microcomputer Applications, 2015, 31(2): 35-38.

[6] Ji Zhen. Analysis of College Students' Consumption Characteristics Based on Campus Card Data[A]. China Statistical Education Society. 2015 (4th) National College Students Statistical Modeling Competition Papers[C].China Statistical Education Society, 2015:27.

[7] Wang Haitao. Research on log-based association rule analysis method [D]. Nanchang Aviation University, 2018.

[8] Ni De,Ma Chuanxiang.Application of FP-growth algorithm and its optimization in tax system[J].Computer Applications,2018,38(S2):140-143.

[9] Wu Shuang. Frequent Itemsets Mining Algorithm and Its Optimization [D]. Huazhong University of Science and Technology, 2018.

[10] Wang Ting. Research and application of frequent item set mining algorithm [D]. North University of China, 2019.

[11] Jiawei Han,Jian Pei,Yiwen Yin. Mining frequent patterns without candidate generation[J]. ACM SIGMOD Record,2000,29(2).

[12] Lu Xianguang, Du Xuehui, Wang Wenjuan.Alarm Correlation Algorithm Based on Improved FP Growth[J].Computer Science,2019,46(08):64-70.