The Research of Mining Association Rules between Personality and Behavior of Learner under Web-Based Learning Environment

Du Jin[1], Zheng Qinghua[1], Li Haifei[2], Yuan Wenbin[1]

[1] Department of Computer Science, Xi'an Jiaotong University, Xi'an, Shannxi, P. R. China, 710049
[2]Department of Mathematics and Computer Science at Union University, Jackson, TN, U.S.A
Xian_dj@163.com

Abstract: Discovering the relationship between behavior and personality of learner in the web-based learning environment is a key to guide learners in the learning process. This paper proposes a new concept called personality mining to find the "deep" personality through the observed data about the behavior. First, a learner model which includes personality model and behavior model is proposed. Second, we have designed and implemented an improved algorithm, which is based on Apriori algorithm widely used in market basket analysis, to identify the relationship. Third, we have discussed various issues like constructing the learner model, unifying the value domain of heterogeneous model attributes, and improving Apriori algorithm with decision domain. Experiment result indicated that this algorithm for mining association rules between behavior and personality is feasible and efficient. The algorithm has been used in a web-based learning environment developed at Xi'an Jiaotong University.

1 Introduction

Web learning is a very promising area and it has the potential to revolutionize the education industry. Today, more than 420 universities have set up web-based virtual school in American， the number of student enrolled is about 1 800 000 and the kind of curriculums reaches 50 000, which overcover almost all of subjects of American university. Moreover, web-based learning has been applied in staff continued training among 60% of US enterprises.

Discovering the relationship between personalities and behaviors of leaner is an important issue in web-based learning environment. It's known that the personality of learner will affects his or her learning behavior mode to a certain extent. On the other hand, some kind of behaviors sequence must be the representation of some personality. However, although the psychologists and pedagogues had been studying the relations between personality and behaviors of learners for many decades, the web-based learning technology is a novel mode of modern education and it is new research field which combined with the psychology, cognitive science, information science, computer technology and so on. So, the problems to discovery the relationship between personality and behaviors focus on the two problems.

The first one is how to build the learner model, it is the key step to get the personality of learners in web based leaning system. Usually, the pedagogues study the learners model under the traditional learning environment, but research of learner model in web based learning system is still a blank field which have profound study foreground. Because the raw data have some specialty of their own, which are complex, dynamic, distributed, so the feature of the learners can not be represented by raw data directly. The complexity was demonstrated in that the dimension of attributes which describe the learner's personality is various, and the value type of the attributes is the quantitative, moreover, the learning behavior model is heterogeneous; The dynamic means that the data resource (e.g. web log on studying) was updated frequently; The distributed denotes that the resource of the data is broad, that include various questionnaire for personality and huge study logs in detail. So the learner model should be represented by mathematical model which is the combination of various science and technology such as information science, pedagogics and so on.

The second problem is which method should be selected to discovery the relationship between the personality and behavior and which is more adaptable and efficient? According to the features of learner model and the demand we expected, the technology of data mining is adopted. But the general algorithm to find the association rules between the 'behavior-personality' for learners have some shortage such as low efficiency, heavy calculation and redundancy results. In this paper, aiming to the particularity of the data to be mined, algorithm with decision attribute based on Apriori is proposed, the algorithm based on the personality model and behaviors model, and the advantage of algorithm was proved in application.

The following are the contributions of this paper. First, this paper proposes the personality mining as one of the key components in personalized web learning environment. Second, Apriori algorithm, which has been widely used for market basket analysis, has been extended with Decision Domains for personalized web learning environment. Third, an extensive experiment has been conducted to validate the feasibility of the new algorithm. As far as we know, our paper is the first to address the issue of relationship between personality and behavior for web-based learners. The rest of the paper is organized as follows. Section 2 discusses Learner Model, which include both personality model and behavior model. Section 3 describes the algorithm; Section 4 describes the performance evaluation. Section 5 describes the possible applications in personalized web-based learning. Section 6 describes the experiments. Section 7 concludes the paper with future work.

2. Learner Model

The learner model is the basis of personality network learning system. Today, the IEEE LTCS has proposed the IEEE 1484.2 PAPI (Public and Private Information), which describe the information about the learners such as age, background, region and so on, and include demographics, major, management, relation, security, preference, performance, works and so on information of learner at eight aspects[1].

Although the PAPI can content the demand of education well, under the personalized network learning environment, it is incapable to be taken as the basis to making out the strategy for personalized education yet. According to the demand on the personalized learning, we enlarge the learner model at non-intelligent aspects by introducing five factors such as personality, motive, concept, method. Furthermore, we import six kinds of study behavior in network education, such as courseware learning, web-based homework/examination, posting and browsing on BBS, answer question by web and interaction among teachers and students by multi-model. We constructed the learner model for the application of personality education. As the figure 1 shown, the grey blocks are the content what should be expanded.

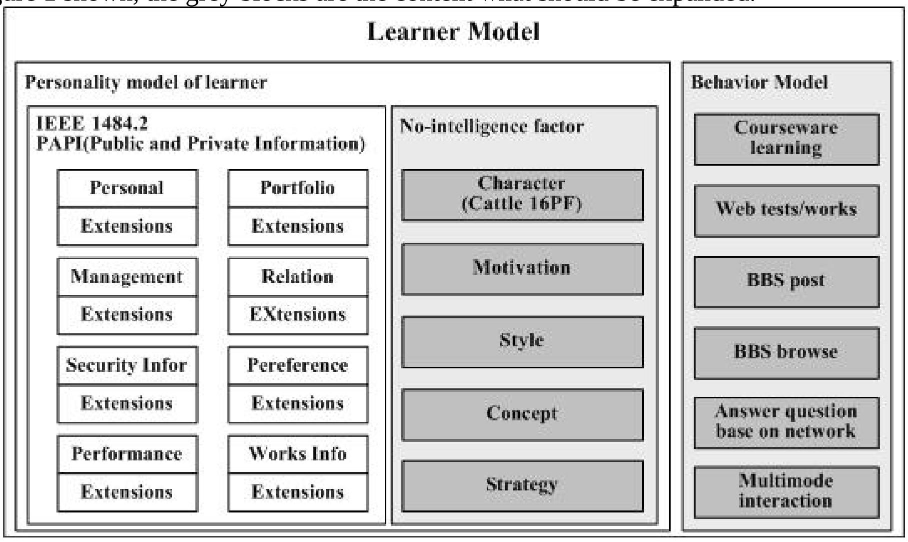

Fig. 1. The structure of LM (Leaner Model)

As the fig.1 shown, the Learner Model include both static model (described with personality model) and dynamic model (described with behavior model), it should be defined as the following tuple:

$$LM = < PM, BM > \quad (1)$$

2.1 Personality Model

PM was composed of some sub-model such as *character, motivation, style, concept* and *strategy,* which are stable elements in LM and represent the various profile of learner in personality respectively. The PM can be defined as:

$$PM = < L\_P, L\_M, L\_S, L\_C, L\_T > \quad (2)$$

Character was the summation of mentality tendency, which is relatively stable for some ones[3]. It was proved that character affected not only the style of interaction in common life, but also the style of behavior in learning[4]. In this paper, the personality was described by means of Cattell's 16PF (Personality Factor).

The **16PF(L_P)** was defined as following tuple:

$$L\_P ::= < UID, A, B, C, E, F, G, E, F, H, I, L, M, N, O, Q_1, Q_2, Q_3, Q_4 > \quad (3)$$

The UID is the unique identity of the learner, the other components are the 16 personal factors, as the table 1 shown, each factor can be measured on a scale, determined by completing a questionnaire, and the word pairs below indicate the extremes of each scale. The letter codes were ascribed to each scale as a shorthand notation.

Table 1. Cattell's 16 Personality Factors

Factor		Descriptors	
A	Warmth	Reserved	Outgoing
B	Reasoning	Less Intelligent	More Intelligent
C	Emotional Stability	Affected by feelings	Emotionally stable
E	Dominance	Humble	Assertive
F	Liveliness	Sober	Happy-go-lucky
G	Rule Consciousness	Expedient	Conscientious
H	Social Boldness	Shy	Venturesome
I	Sensitivity	Tough-minded	Tender-minded
L	Vigilance	Trusting	Suspicious
M	Abstractedness	Practical	Imaginative
N	Privateness	Straightforward	Shrewd
O	Apprehension	Self-Assured	Apprehensive
Q1	Openness to Change	Conservative	Experimenting
Q2	Self-Reliance	Group-dependent	Self-sufficient
Q3	Perfectionism	Self-conflict	Self-control
Q4	Tension	Relaxed	Tense

Similarly, the other elements of PM can be defined as following formal expression.

Study Motivation Information of Learner:

$$L\_M ::= < UID, M\_C, M\_I, M\_R, M\_D, M\_S, M\_E > \quad (4)$$

Where M_C is information of challenge, M_I is interest, M_R is curiosity, M_D is independence, M_S is Success and M_E is extrinsic motivation.

Study Style Information:

$$L\_S ::= < UID, S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8 > \quad (5)$$

The 'Study Style' was defined as the composite of characteristic cognitive, affective, and physiological factors that serve as relatively stable indicators of how a learner perceives, interacts with, and responds to the learning environment (Keefe, 1979). The styles is multiple, Where S1: field independence | field dependency; S2: impulsive | reflectivity; S3: holist | analytical; S4: serial | random; S5: group-oriented | individual-oriented: S6: visual | auditory | hand-on; S7: objective | nonobjective; S8: open | close.

Study Concept Information:

$$\text{L\_C} ::= < UID, C\_M, C\_E, C\_A > \quad (\mathbf{6})$$

Where C_M is Self-Management, C_E is Self-Efficiency, C_A is Self-Attribute.

Study strategy Information:

$$\text{L\_T} ::= < UID, T\_M, T\_C, T\_E, T\_R, T\_S, T\_O > \quad (\mathbf{7})$$

Where T_M is memory, T_C is cognize, T_E is self-management, T_R is retrieve, T_S is sense and T_O is intercommunicate.

In order to get the value of those attributes, we designed the various questionnaires respectively to collect the information of every element in PM. For example, concerning Cattell's 16PF, we us the classic questionnaire with 189 questions to get all values of 16 personality factors.

2.2 Behavior Model

The learning behaviors are action sequences of learner recorded by network learning log, and what should include all information of learner's activity under different learning mode which provided by learning system. Here, we describe the BM from 6 different aspects:

$$BM = \{B\_C, B\_T, B\_B1, B\_B2, B\_A, B\_I\} \quad (\mathbf{8})$$

B_C (Behavior on courseware learning): <User id, course id, unit id, entry, stay time>

B_T(Behavior on Test/homework):<User id, course id, test/homework id, finished, score, question sum, correct ratio>

B_B1(BBS posting): <User id, course id, article id, article type (post, reply), words, quality>

B_B2(BBS browsing): <User id, course id, article id, stay time>

B_A(Behavior of answer question): <user id, course id, speaks, speech words, online time>

B_I(Behavior of interaction): <User id, course id, class id, type(multimedia, e-board, text chat, application share), time>

Generally, the value of each attribute was gained statistically, and the BM represents the profile of the learner's behavior quantificationally. However, the different attribute has the different value field and type, in order to content the need of following mining, we should map the value of attributes of PM and BM into uniform integer range in dividing partition. For example, we can mapped the quantitative value into three interval: *high, middle, low* partition, and represent them with ascent integer *1,2,3*. After data converted, the every attributes of the LM (Learner Model) have the same value type and the same value range (e.g. 1,2,3). According to the type of attributes, the association rule mining among behavior and personality characters can be performed.

3. Algorithm to Mine association rules between behavior and personality

3.1 problem explanation

After mapping the value of behavior model and personality model into uniform range, we can use the tradition algorithm Apriori to analyze the association rules among the attributes of behaviors and personalities. Firstly, let us define the problem more clearly use the following mathematical model[2]

Definition 1: association rule

- $I = \{ i_1, i_2, ..., i_m \}$ the set of items
- Database *D* is a set of transactions.
- Transaction *T* is a set of items such that $T \subseteq L$.An unique identifier, *TID*, is associated with each transaction.
- *T* contains *X*, a set of some items in *L*, if $X \subseteq T$.
- Rule form: “Body ⇒ Head [support, confidence]”
- Association rule, $X \Rightarrow Y$ $X \subset T$, $Y \subset T$, $X \cap Y = \varnothing$
- Confidence – % of transactions which contain *X* which also contain *Y*.
- Support - % of transactions in *D* which contain $X \cup Y$.

In our problem, the items are all attributes with values; the database D is the all records which include the results of questionnaires and of the log analysis on behavior sequence. The transaction T presents the description for one user, the user id is as same as the *TID*, which joint the character vectors of PM and BM together, so the all attributes can be seemed as undifferentiated each other. At last, we set the special threshold（min-support and min-confidence） to get the association rules between the PM and BM.

However, the traditional algorithm, as Apriori or Aprioritid[6], deal with all attributes without distinguish. There are 3 types of rule mode as 1) PM-PM, 2) BM-BM, 3)Mixed. Especially, the last type is very complex, the body and head of rules maybe include the attributes of PM or BM or both. In order to get the rules as we expected, we have to scan all records to find all potential rules, and delete the rules uninterested. So some cost of time and calculation was wasted on analyzing and deleting phases.

In order to improve the efficiency of algorithm, and avoid the unnecessary cost, we should modify the traditional algorithm according to the demand of application. In our problem, we use the behavior sequence as the input parameters, and generate the rules as BM⇒PM. The elements of rule body belong to the attribute set in BM and elements of rule head belong to the PM's. In this way, we can deduce the personality from the learner's behavior sequence recorded by web logs.

Well, we can divide the all the items in Learner Model into different fields, and through the association analysis, to discover the relationships among the items (attributes with value) which belong to PM and BM. So, let's make definition at first.

Definition 2: domain

The attributes set I in LM can been divided between 2 subsets: $I = I_P \cup I_B$, and $I_P \cap I_B = \phi$. We named I_P, I_B as a domain, and the domain I_P, I_B can been expressed as $I_P = \{p_1, p_2, \cdots, p_n\}$, $I_B = \{b_1, b_2, \cdots, b_m\}$.

In this paper, it proposes an 'Algorithm of Apriori with Decision Domains (named as DD for short)'. This algorithm can discover the rules of which the structure was foreknown (the items of head and body belong to different domains respectively, and the domain of head was expected).

3.2. Algorithm Analysis

Take a fact about association analysis into account.

If $a_1,...,a_i \Rightarrow b_1,...,b_j$ existed, $a_1,...,a_i \Rightarrow b_1$, $a_1,...,a_i \Rightarrow b_2$,…, $a_1,...,a_i \Rightarrow b_j$ must existed too, any subset of a frequent item set must be frequent. So, we can translate the problem $I_A \Rightarrow I_B$ into the rule set as $\{\wedge a_{i'} \Rightarrow b_{j'}\}$.

Definition 3: Decision Domain (DD)

Supposing the structure of the rules and the head of rules are foreknown, we want to find out the associate rules such as' $b_i, b_j, ..., b_m \Rightarrow p$ '.where, $\{b_i, b_j, ..., b_m, p\}$ is attributes set, $b_i, b_j, ..., b_m$, belong to domain I_B, p belong to I_P. Here, the head of rule p was known and name as *Decision Domain(DD for short).*

Meanwhile, we notice such facts:

Theorem 1

During the association rules mining with DD, if k-items ($b_1, b_2, ..., b_{k-1}, p$) (the length of attribute tuple is k) is not a frequent items set, according to the Monotonicity Property of frequent items (A subset of a frequent itemset must also be a frequent itemset), the $b_1, b_2, ..., b_{k-1}$ must been invalid frequent items to generate the rule as $b_1, b_2, ..., b_{k-1}, ..., b_n \Rightarrow p$.

In this theorem, the 'invalid' means: even though ($b_1, b_2, ..., _{k-1}, p$) can generate the frequent items such as ($b_1, b_2, ..., b_{K-1}, ..., b_n$), it can not generate frequent itemset as ($b_1, b_2, ..., b_{k-1}, ..., b_n, p$).

Therefore, while generating the association rules, we can delete the frequent items such as $b_a, b_b, ..., b_i, b_j, ..., b_n$, which only generate the rules as $b_a, b_b, ..., b_i \Rightarrow b_j, ..., b_n$,from all set of frequent itemset.

The process of 'Algorithm of Apriori with DD' is as following:

Step1: divide L_k (k-frequent items) between L_{ki} which includes decision domain and L_{k2} which excludes decision domain, both L_{k1} and L_{k2} are k-frequent items.

Step2: generate the k-candidate set $C_{(k+1)1}$ which includes decision domain from L_{k1} 、 L_{k2}

Step3: counting the items in $C_{(k+1)1}$, generate (k+1) frequent items $L_{(k+1)1}$ which include decision domain.

Step4: supposing the item which included in $C_{(k+1)1}$ and excluded in $L_{(k+1)1}$ is $b_i, b_j, ..., b_k, p$;

Step5: delete all of items which include $b_i, b_j, ...b_k$ from L_{k2}

Step6: generate k+1 candidate $C_{(k+1)2}$ which exclude decision domain from L_{k2}

Step7: counting the items in $C_{(k+1)2}$, generate k+1 frequent items $L_{(k+1)2}$ which exclude decision domain;

Step8: repeats step 1 ~ step 7 till the largest set of frequent items is generated.

Example:

Table 2. Transaction data

UID	Items list
U_1	I_1， P
U_2	I_2， I_4， I_5
U_3	I_1， I_2， I_4， P
U_4	I_1， I_2， I_4， I_5
U_5	I_2， I_4， P
U_6	I_2， I_4， I_5
U_7	I_1， I_2， I_4， I_5， P
U_8	I_2， I_4， P
U_9	I_2， I_5， P
U_{10}	I_2， I_3， I_4， P

As the table 2 shown, I_1, I_2, I_3, I_4, I_5 are feature attributes of BM, P is the character of PM, totally 10 transactions in database D, |D|=10. We need to get the rules form as $I_i, I_j, ..., I_n \Rightarrow P$ through mining.

Set the min-support = 30%, the count of min-support is 3.

If we use the traditional algorithm such as Apriori, then we achieve the analyed result as table 3:

If we adopt the algorithm with DD, to delete the invalid frequent items set in course of mining, the result should be shown as table 4.

Table 3. The course of Traditional Apriori analysis

C_1	
Itemset	Supp
I_1	4
I_2	9
I_3	1
I_4	8
I_5	5
P	7

L_1
Itemset
I_1
I_2
I_4
I_5
P

C_2	
Itemset	Supp
I_1, I_2	3
I_1, I_4	3
I_1, I_5	1
I_1, P	3
I_2, I_4	8
I_2, I_5	5
I_2, P	6
I_4, I_5	4
I_4, P	5
I_5, P	2

L_2
Itemset
I_1, I_2
I_1, I_4
I_1, P
I_2, I_4
I_2, I_5
I_2, P
I_4, I_5
I_4, P

C_3	
Itemset	Supp
I_1, I_2, I_4	3
I_1, I_2, P	2
I_1, I_4, P	2
I_2, I_4, I_5	4
I_2, I_4, P	5

L_3
Itemset
I_1, I_2, I_4
I_2, I_4, I_5
I_2, I_4, P

Table 4. The course of association analysis with decision domain

C_{11}	
Itemset	Supp
P	7
C_{12}	
I_1	4
I_2	9
I_3	1
I_4	8
I_5	5

L_{11}
Itemset
P
L_{12}
I_1
I_2
I_4
I_5

C_{21}	
Itemset	Supp
I_1, P	3
I_2, P	6
I_4, P	5
I_5, P	2
C_{22}	
I_1, I_2	3
I_1, I_4	3
I_2, I_4	8

L_{21}
Itemset
I_1, P
I_2, P
I_4, P
L_{22}
I_1, I_2
I_1, I_4
I_2, I_4

C_{31}	
Itemset	Supp
I_1, I_2, P	2
I_1, I_4, P	2
I_2, I_4, P	5
C_{32}	

L_{31}
Itemset
I_2, I_4, P

In table4, when L_{11}、 L_{12} generate C_{21}、 C_{22}, the following step needed to be executed.

Step1: self-joining L_{11}, (as $L_{11} \triangleright\triangleleft L_{11}$). For the L_{11} have one 1-item only, so it can not generate 2-items.

Step2: join L_{11} with L_{12}(as $L_{11} \triangleright\triangleleft L_{12}$), generate (I_1, P)、(I_2, P)、(I_4, P)、(I_5, P) 2-candidate items. After count supports of the 2-candidates, the support of (I_5,P) is under the min-support, so (I_5,P) is deleted. According to theorem 1, I_5 is invalid frequent item, so it should be deleted from L_{12}.

Step3: self-joining L_{12}(as $L_{12} \triangleright\triangleleft L_{12}$),in which the I_5 had been deleted, counting supports of the three 2-candidates(I_1,I_2), (I_1,I_4), (I_2,I_4) respectively;

Step4: generate L_{21} and L_{22} from C_{21} and C_{22};

From the table4, it is known that, the association analysis algorithm with decision domain can descend the complexity of association analysis efficiently, through removing invalid frequent items and reducing the pass of database scanning,

4. Time Complexity Evaluation

In this section, we will compare time complexity of algorithm optimized with that of traditional Apriori algorithm, to evaluate the performance and show the advantage of the algorithm proposed.

Considering the complexity of algorithm is affected by the concrete transactions sequence, in order to illuminate more clearly, we make several suggestion first of all.

On the assumption that length of LM (the number of attributes of PM and of BM) in this database D is *m*, the sum of items belong to body domain(BM) is *m-1*, the item belong to head domain(PM) is *1*. There are *n* transactions in database D. For the Apriori-generate C_2 from L_1 is the key step in all algorithm process, so, in this paper, we evaluate the advantage of this algorithm according to the sum of attributes which can be removed from L_1.

Supposing, when generate L_1 from C_1, the sum of items can be removed no-frequent items is *M*, the sum of invalid frequent items is *N*, so there are:

The time complexity of traditional algorithm Apriori would be illuminate as:

$$O(n*(C_m^1 + C_{m-M}^2 + ... + C_{m-M}^{m-M-1})) = O(n*2^{m-M}) \tag{9}$$

The time complexity of Apriori Algorithm with Decision Domain is illuminated as:

$$O(n*(C_m^1 + C_{m-M-N}^2 + ... + C_{m-M-N}^{m-M-N-1})) = O(n*2^{m-M-N}) \tag{10}$$

Obviously, when $N \geq 1$, then

$$O(n*2^{m-M}) >> O(n*2^{m-M-N}) \tag{11}$$

Hence, the algorithm with decision domain can reduce the number of k-candidates C_k efficiently, moreover, descend the time complexity in association rules analyzing.

5. Application

Under the web based learning environment, we describe the character of learners in two aspects, Behavior Model and Personality Model, the former was expressed by a set of vector such as $\vec{B} = (B_1, B_2, ..., B_m)$. The latter is static, gained through several questionnaires, such as Cattell's 16PF, can be expressed as:

$$\vec{p}_{L\_K} = (A, B, C, E, F, G, E, F, H, I, L, M, N, O, Q_1, Q_2, Q_3, Q_4) \tag{12}$$

We need to explain what does A, B, C, up to Q4 mean. Otherwise, people will have hard time to understand it.

We will get the rule such as: $\bigwedge_{b \in B', B' \in B} ((b, s'), s' \in S) \rightarrow (p, v')$ by our algorithm. B is the set of behavior attributes. B' is a subset of B and is not empty; S is the value range of some attribute b in B', s' is a possible value in S; $p \in P$, v' is a possible value in p.

For the limited of application condition, we analyzed only the behaviors of BBS, B_C and B_A to discover the relationship between character attributes and behavior. We set the min-support is 10%, and the min-confidence is 60%, the results of rules searching as the table shown.

6. Experiments

We developed 'Personalized English Learning System' and applied it in Xi'an Jiaotong University. After one month's using, we collected 324 students Cattell characters and about 146 000 web log records. The experiment show that, when the structure of rules is constrained, by dividing items into different domain, and filtering the items in the course of mining, the algorithm with decision domain will avoid the generation of redundancy rules, reduce the complexity of calculate and improve the analysis efficiency.

Table 5. the result compared between Apriori and Apriori with decision domain

Character	Apriori	Apriori with DD
amount of transaction examples	324	324
Amount of available rules	67	29
Running time	1420ms	843ms

According to the results of association analysis, the relationship between PM and BM as the table7 shown: :

Table 6. part of relationship between BM and PM

BM / Personality	B1	B2	B3	B4	B5	B6
A(Warmth)	PC			PC		PC
E(Dominance)		PC	PC	NC	PC	
G(Rule Consciousness)	PC					
L(Vigilance)	PC		NC		PC	

Q2(Self-Reliance)		PC	PC			
……						

Where, NC: negative correlation, PC: positive correlation, B1: Stay time in B_C, B2: The proportion of article type in BBS, B3: The words in B_BBS1, B4: The quality of article, B5: The proportion of question answered, B6: Total online time in B_A.

7. Conclusions and Future Work

The experiment show that, the Apriori algorithm with DD can discover the relationship between personality and behavior, improve the efficiency of mining. After get the personality features of learner by analyzing his/her behavior, how to adopt the proper study strategy and settle adaptive leaning material, will be our future work.

Reference

1. IEEE Learning Technology Standards Committee (LTSC), IEEE 1484.2 “PAPI Learner Model”
2. Agrawal R, Srikant R. Fast algorithm for mining association rules. Proceedings of 1994 International Conference on Very Large Databases. Santiago, Chile, 1994, 487-499
3. E. Vance Wilson, Student characteristics and computer-mediated communication, Computers & Education, 2000(34), 67-76
4. Eyong B. Kim, The role of personality in Web-based distance education courses Communications of the ACM, Volume 47, Issue 3 March 2004
5. Carey, J. M., Kacmar, C. The Impact of Communication Mode and Task Complexity on Small Group Performance and Member Satisfaction. Computers in Human Behavior, 13(1), 23-49
6. Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, Morgan kaufamnn Publishers. 2001
7. Liu Jun, Li Renhou, Zheng Qinghua, Study on the Personality Mining Method for Learners in Network Learning, Academic Journal of Xi'an Jiaotong University, 38(6)，2004. (EI indexed, AN: 04358329828)