An Attributes Correlation Based Learning Guidance Model

Binbin HE, Zhiyi FANG, Bo SUN, Lingxi ZHOU and Yunchun ZHANG
College of Computer Science and Technology, Jilin University,
Changchun, 130012, P.R.China
hebinbin1989@gmail.com, zyfang@public.cc.jl.cn

Abstract
Using the theory of Web Data Mining, this paper proposes a physics discipline personalized learning evaluation system model based on decision tree algorithm. The overall structure of this system model, which is applied to the network-based education, is presented. The key part of the model, data collection module and personalized evaluation module are introduced. The advantages and disadvantages of the ID3 algorithm and the C4.5 algorithm are analyzed respectively. Taking the comparison of the two algorithms and the unique characteristics of the personalized learning system into consideration, a new C4.5r decision tree rule simplification algorithm based on attributes correlation is proposed in this paper. The results from a set of experiments shows that the new C4.5r decision tree algorithm outperforms the previous ones in running time, the size of the rule sets and overhead.

Keywords: Web data mining, personalized learning, decision tree algorithm.

1. INTRODUCTION
Nowadays, network education is mostly in the state of sharing teaching resources. Patterns of traditional classroom-based education have changed dramatically. With the emergence of network and computer, a network-based education called E-learning can do the same or even more things as traditional education dose. Although anybody authorized can access the teaching resources, the number of people who can benefit from the network-based education doesn’t increase correspondingly, as the current teaching system pays little attention to users. It can neither provide appropriate, personalized and interactive learning environment according to learner's level and learning situation, nor automatically track the learner's interests to provide the personalized content. By taking advantage of the knowledge acquired from the analysis of the user's behaviour as well as other information such as structure, content and user profile data, Web personalization customizes a Web site to the needs of specific users [1]. Personalization has become a reality and is possible by using efficient methods of data mining and knowledge discovery [2]. Focuses have been put on developing e-learning systems with personalized learning mechanisms to assist on-line Web-based learning and to adaptively provide learning paths [3]. Most of the research in this area mainly focuses on resource recommendation, while overlooked the overall requirement of different users. Based on the current development in both data mining and personalized education system, the users can benefit more than before. In this paper, a learning guidance model based on Data Mining technology in a network environment is proposed and implemented. This model will apply data mining technology to the excavation of the student learning information database, try to discover the weakness in the student learning process, and then abstract the strategy which might guide the learning process efficiently.

This paper is organized as following. First of all, in section 2, it is necessary to give a preview about data mining technology. In section 3, the learning guidance model is presented in detail. In section 4, the mining algorithm used in this paper and its derivation are introduced. Finally, the experimental results and conclusions are listed in section 5 and 6, respectively.

2. WEB DATA MINING
Web Data Mining is an application of Data Mining technology under the web environment. It is a model that can excavate implicit, unknown and special patterns which have potential values. As the diversity of the web varies the duty of Web Mining, Web Mining can be divided into three parts: Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining is regarded as the combination of Web information retrieval and web information extraction. Web Structure Mining is usually used for mining the hyperlink structure of the Web page in order to discover the information included in the hypertext structure. Web Usage mining is a technology which can predict a user’s browsing behaviour by mining the information from the Web server's log files.

3. MODEL DESIGN
3.1 The Improvement of the Traditional Network Teaching Environment
The structure of the current network-based education platform, as shown in Figure 1, is generally composed of three parts: Teaching Resources Library, Learning Platform and Users.
The personalized network-based teaching platform not only improved original functions, but also introduced personalized analysis and evaluation module into the system. As shown in Figure 2, after login and status confirmation, the information-gathering module of this system begins to collect the user’s requests and track the user’s behaviour, then sends the information collected to the database, processes the personalized analysis and evaluation, and finally sends back the analysis results to the user. This set of processes gives the platform the ability to provide specific teaching resources according to the personal characteristics of each student. Although this system concentrates on providing personalized learning for those who study physics, it is not constrained. Thus, it can also be applied in different teaching domains without great modification.

### 3.2 Data Collection Module

The personality of each user is mostly abstracted from the sheer volume of data collected previously. The precision of the learning guidance highly depends on how accurate the system can recognize the personality of each user. Data collection module is mainly responsible for collecting data from students’ learning behaviour online and storing them into the database. In conclusion, this module is the basis of the entire physical personalized teaching platform. As it is also the data source of the personalized analysis engine, the quality and quantity of the information collected will directly affect the accuracy of personalized system analysis.

As shown in Figure 3, the system collects the information from multiple sources, including written materials, knowledge study, online tests and so on. The preference and the whole learning process are also recorded for the subsequent mining. The information collected from different sources should be correlated to meet the requirement of getting an overall view about each user. By doing this, the learning guidance can be achieved.

### 3.3 The Design of Physics Discipline Personalized Evaluation Module

The main task of personalized evaluation module is to select an appropriate Data Mining algorithm, to establish a personalized learning effect model, and to evaluate the student’s learning online in physics. The model of Physics discipline personalized evaluation module shows in Figure 4.
4. **ALGORITHM DESIGN**

4.1 **Comparisons of ID3 Algorithm and C4.5 Algorithm**

In the process of constructing a decision tree, the Information Gaining method is usually used to help generate branches according to the values of different attributes. Suppose S as a set of samples whose attribute information has been calculated. Next, select the attribute with the largest gain as the test attribute of the given set S, and then produce the corresponding branch nodes [4][5]. Step by step, ID3 [6] refines the decision tree until a completely correct decision tree through the constant cycle treatment is found, and finally conduct a top-down induction to form a group of rules in the format of IF ......THEN.

In summary, the ID3 algorithm has two major advantages. Firstly, the objective function is limited to a supposed search space, and there is no risk of having no solution. Secondly, the training data is entirely absorbed, which means that the statistical character of the whole training examples will be taken into account when making the final decisions, in order to resist the noise. Nevertheless, the ID3 algorithm has shortcomings as well. First of all, only one solution is maintained in the process of searching. In addition, there is no backtracking while do

search, and the algorithm may converge to a partial optimal solution rather than the global optimal solution. At the same time, it is a kind of decision tree algorithm with only a single variable, which means that it is difficult to express complex concepts.

The C4.5 [7] is a decision tree generating algorithm developed from the basis of the ID3. This decision tree algorithm can be divided into three parts [8] according to the basic principles of C4.5 algorithm: decision tree generation algorithm (C4.5tree), pruning algorithm (C4.5pruning) and rules generation algorithm (C4.5rules). Compared with ID3 algorithm, C4.5 algorithm adds the capability of processing continuous attributes and the situation where property values absence. This greatly improves the efficiency of the algorithm. C4.5 algorithm is widely used because of its quick classification and high precision. In this system, the majority of data samples used in data mining have continuous attribute values. Therefore, C4.5 algorithm is selected as the algorithm carried out by the learning behaviour evaluation module.

The following are some other decision tree algorithms which are usually used: CART algorithm [9], SLIQ algorithm [10], and SPRINT algorithm [11].

4.2 **Improvement of Algorithm**

When applying C4.5 algorithm to classify some unknown samples, the system may encounter the "over-fitting" problem. As a result, it is necessary to simplify the samples before they are classified.

The following are the processes of the improved algorithm, named new_C4.5r:

1. Use C4.5tree to construct a complete decision tree T.
2. T will be converted to the rule set R. The rule r corresponds with a path from the root node to a leaf node in the T.
3. R: \{r_i \} if Cond_1 ∧ Cond_2 ∧ ... ∧ Cond_n then class C_i \}.
4. Simplify each rule r_i of R as following:
   - i = 1;
   - While (i ≤ n)
     - \( t_{i+1} = P(\text{Cond}_{i+1} \land \text{Cond}) \);
     - if \( t_{i+1} ≥ \lambda \) then
       - \( i = i + 1 \);
     - else
       - delete \( \text{Cond}_{i+1} \land ... \land \text{Cond}_n \) in r;
       - break;

Figure 4: Model of evaluation module
In the above process, introduce the parameter $\lambda$ as the threshold value of $P(\text{Cond}_i \land \text{Cond}_j)$. It is similar to the concept of the Minimum Support in the Association Rules. Its default value is 0.15% [12]. The value of $\lambda$ is controlled to eliminate the over-fitting part of the rules. Merge and simplify the same rules in $R$, and get a new rule set $R'$.

Establish an attribute-associated matrix $(T_{vs})_{n \times n}$:

If $t_{vs} = 0$, Attribute A and Attribute B are irrelevant.

If $t_{vs} = 1$, Attribute A and Attribute B are relevant.

Simplify each rule $r'$ of $R'$ as following:

$i = 1$;

while ($i \leq n$)

{ 
for ($j = i + 1, j < n, j++$)

{ 
if $t_{vs} = 0$ then continue;
if $t_{vs} = 1$ then 

Calculate the condition probability in the training set:

$\text{P}(\text{Cond}_j | \text{Cond}_i)$, $\text{P}(\text{Cond}_i | \text{Cond}_j)$;

if $\text{P}(\text{Cond}_j | \text{Cond}_i) \geq \text{P}(\text{Cond}_i | \text{Cond}_j)$ then

Tag and eliminate Condj;

if $\text{P}(\text{Cond}_j | \text{Cond}_i) < \text{P}(\text{Cond}_i | \text{Cond}_j)$ then

Tag and eliminate Condi;
break;
}

$i = i + 1$;
while (Condj is tagged)

} 

In this process, at first, the correlation between attribute $A_i$ and $A_s$, which are belong to $\text{Cond}_i$ and $\text{Cond}_j$, respectively, is judged. If $A_i$ relates to $A_s$, $\text{Cond}_i$ and $\text{Cond}_j$ in rules will be kept, otherwise, both conditional probabilities are calculated, and $\text{Cond}_i$ or $\text{Cond}_j$ will be eliminated according to the confidence. The same rules of $R'$ are merged and simplified, and a new rules set $R''$ is obtained [13].

5. EXPERIMENTAL RESULT ANALYSIS

At first, by using C4.5tree algorithm, the complete decision tree was structured with the data in test data set. And then two rule sets were produced based on the C4.5rules algorithm and the new_C4.5r algorithm. The data in Table 1 are the mean values of the results from ten experiments. The size of the rule set, the number of rules, running time and classification error rate of two algorithms are compared under the circumstance that the size of training set is fixed.

In Table 1, during extracting and simplifying of complete decision tree rules, the classification error rate of the new_C4.5r algorithm is close to that of the C4.5rules algorithm. The size of the rule set and the number of rules are smaller than those of the C4.5rules algorithm. Therefore, when extracting and simplifying the complete decision tree rules, the systems with improved algorithms can not only maintain the classification precision, but also produce simpler rule set and improve the speed and efficiency of the computation.

<table>
<thead>
<tr>
<th></th>
<th>C4.5rules algorithm</th>
<th>new_C4.5r algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>the size of training set</td>
<td>20000</td>
<td>20000</td>
</tr>
<tr>
<td>running time (S)</td>
<td>73</td>
<td>42</td>
</tr>
<tr>
<td>classified right (%)</td>
<td>89.2</td>
<td>87.7</td>
</tr>
<tr>
<td>the number of rules</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>the size of the rule set</td>
<td>195</td>
<td>113</td>
</tr>
</tbody>
</table>
6. CONCLUSIONS

On the basis of Web Data Mining theory and the current development in personalized educational system, this paper establishes a model of the Physics discipline personalized learning evaluation system using decision tree algorithm. The advantages and disadvantages of ID3 algorithm as well as C4.5 algorithm are analyzed and compared. Then a new C4.5r decision tree rule simplification algorithm based on attribute correlation is proposed. After applying it to the personalized learning evaluation system, the experimental results proved that the new C4.5r algorithm outperformed the traditional C4.5 algorithm in running time, the size of the rule sets and overhead. On future, more functions will be included in this system by using heterogeneous data mining algorithms. Meanwhile, more parameters, not only running time, the number of rules et al., should be considered to give a scientific evaluation to this system.

7. REFERENCES