doi:10.21311/002.31.6.10

Application of Data Mining in Web-based Education Platform

Chao Luo

Neijiang Normal University, Neijing 641100, China

E-mail: luochao85821@163.com

Abstract

This paper discusses the application of data mining on all aspects of the life-long education platform. Firstly, this paper introduces knowledge and common technology of data mining. Secondly, it introduces the application of data mining technology on course recommendation in the life-long education platform. Thirdly, it introduces the application of data mining technology on score analysis in the life-long education platform. Lastly, it introduces the application of data mining technology on assisting teacher’s decisions in the life-long education platform. It is beneficial to introduce data mining technology into web-based education. The data mining technology covers a wide area. This paper will discuss the application of data mining on web-based education from the aforementioned three aspects. With the development of data mining technology and educational information, this paper believes that more and updated technology will be applied to web-based education to make web-based education more intelligent and humanized.

Keywords: Data mining, web-based education platform, education model, education method.

1. INTRODUCTION

In the traditional classroom setting, the teacher can evaluate a learner through tests, homework, classroom learning, answering questions and classroom performance. In the process of web-based education, a learner may conduct humanized learning depending on a web-based teaching system through a computer and the internet. Due to randomness of learning time and modes on the internet and the universality of the learner’s resources, it is difficult for teachers to determine the learning condition and personal details. Accordingly, it is hard to appropriately evaluate and guide learners. (Nancy, 2010)

Due to lack of supervision of the learning process of the web-based teaching platform and without dynamic integration among basic conditions of learners, teachers can only know the conditions of learners through several databases, while answering questions, examination result and discussions on the forum which are mutually independent. In fact, learning features will be reflected to various links because of their own constraints, making it difficult to identify learning patterns and potential learning features reflected to each link by using simple statistical analysis. This is not helpful for teachers who must integrally evaluate learners and make proper decisions and offer feedback to learners for their improvement (Latchman et al., 1999)

The decision support function module can provide content analysis related to learners for teachers, such as registration information, learning activity analysis, homework analysis and evaluation of the learner, as well as providing relative information analysis of course learning for teachers (Athula and Tamara, 2012).
2. DEFINITION OF DATA MINING

Data mining is a process of collecting unknown and potentially valuable information from many incomplete, noisy, random and fuzzy data.

The essential difference between data mining and traditional data analysis (such as query, statement and online application analysis) is to mine information and find knowledge without any express assumption. The information obtained through data mining is usually previously unknown, effective and available (Ashra, 1999). The previously unknown information refers the fact that the information is unforeseen in advance; namely, data mining is to find information that cannot otherwise be obtained through intuition or even counter-intuition. The more unexpected information is, the more useful it can potentially be. The usefulness of information requires careful checking of mined information before mining. Only by guaranteeing the usefulness of information (or data) can the usefulness of mined information be guaranteed (Segpo and Perth, 2000). Most of all, the information obtained will be made available; namely, the information or knowledge should be useful, practical and realizable for discussion in business or research fields.

There are various classification methods for data mining technology. According to the data mining task, the methods include association rules mining, data classification rule mining, clustering mining, dependency analysis, dependency model discovery, concept description, deviation analysis, trend analysis and model analysis. According to databases mined, the methods include relational database, object-oriented database, special database, temporal database, multi-media database and heterogeneous database. According to the technology used, the methods include artificial neural network, decision tree, genetic algorithm, the principle of neighborhood and visualization. The following sections will describe the clustering mining, data classification rule mining and association rules mining in detail (Asleena, 2001).

3. DATA MINING AND ANALYSIS OF LEARNER’S INFORMATION

Data mining is a decision support process of collecting key data that can assist decision-making after extracting, converting, analyzing and modeling data. Data mining is a method for deep-level data analysis (David and Zippy, 2001). In this section, this paper brings association rules mining of data mining technology into evaluation of learning behavior. Online learning behavior data can be obtained by supervising online learning condition of learners. Then, deep-level analysis of learners can be made by combining basic information about learners, which can help teachers to forming visual and reliable evaluation of learners. This section analyzes learners from two aspects: statistic analysis and association rule analysis (Sheulur, 1998).

In the process of learning on a web-based education platform, a learning model must be established for learners and data related to learning behaviors must be collected in this model. Data collection is an important part of data mining so that whether or not the collected data is complete and accurate significantly affects the quality of assessment of learners’ condition (Gordon, 2014).
Select category property

Choose the training set

The calculation of each attribute information gain rate

Choose the maximum of information gain rate as a root node, and according to its value

Determine whether there is any new properties

Yes

no

To generate the decision tree model

Figure 1. The working process of the decision tree and the generation process

A learner logs into the education platform and enters the learning page after the learner has been authenticated and then chooses which courses to learn. The information reflecting the condition of learners includes: the registration form for basic information of learners, course learning, discussion in forum and online testing (Deryn et al., 2008).

4. DESIGN OF INFORMATION ANALYSIS SYSTEM OF LEARNERS

4.1 Structure of Information Analysis System of Learners

The information analysis system of a learner includes three blocks: data acquisition, statistical analysis and association rule mining. The system’s overall structural chart is as shown in Figure1.
The data acquisition module tracks and records the online learning condition of learners, collecting and quantifying learning information of the learner and storing them in the database (Frank, 1995).

The statistical analysis module extracts and processes data in the database and provides a graphic interface to realize real-time statistical analysis of learning behavior of learners.

The association rule mining module applies the association rule mining method to calculation of related data to find related rules (Susan, 2001).

4.2 Data collection

The information about the learner includes: registration form for basic information of learners, course learning, forum discussions and online testing. Regarding the four aspects, the concerned characteristics are different: attributes that can reflect the basic features of learners is selected from the basic conditions of learners; the length of continuous time of online learning is selected for describing course learning; active degree in the forum is focused in the aspect of forum discussion; score levels after a test is focused in the aspect of online test. These different aspects require different data collection processes.
4.3 Design of Statistical Analysis Module

The statistical analysis module processes the data which is collected by the data module and stored in the database and provides a graphic interface to realize real-time statistical analysis on basic information and online the learning condition of learners (Lindauer and Bonnie, 2002).

Data processing and data handling refers to validity verification of data collected by the data collection module and restoration of incomplete records. Statistical analysis predefines the database inquiry mode, designing a multi-inquiry mode combination and showing the learning condition of learners from multiple aspects.

![Figure 4. Learners' overall situation analysis system structure](image)

According to teaching practical teaching experience, statistical analysis conducted from horizontal and vertical aspects can comprehensively understand the condition of the learner.

The vertical aspect refers to the individual learner, aiming at describing the overall online learning condition and details of the learner. For example, it describes the overall learning curve, a detailed learning schedule and participation in forum discussion and tests. Showing this information to the individual learner is beneficial to realize and master the learning progress and learning pattern in a more comprehensive and objective manner.

### Table 1 S_forum table structure

<table>
<thead>
<tr>
<th>Field name</th>
<th>Character types</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studentno</td>
<td>Varchar</td>
<td>Student id</td>
</tr>
<tr>
<td>Courseid</td>
<td>Int</td>
<td>Course number</td>
</tr>
<tr>
<td>Status</td>
<td>Int</td>
<td>Post type</td>
</tr>
<tr>
<td>Post_time</td>
<td>Numeric</td>
<td>Time of occurrence</td>
</tr>
<tr>
<td>Postid</td>
<td>Int</td>
<td>Post type</td>
</tr>
<tr>
<td>Postcount</td>
<td>Int</td>
<td>The total number of post</td>
</tr>
</tbody>
</table>

### Table 2 Learners' basic information analysis

<table>
<thead>
<tr>
<th>Field name</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studentno</td>
<td>Student id</td>
</tr>
<tr>
<td>Studenttype</td>
<td>The learner types</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender</td>
</tr>
<tr>
<td>Xueli</td>
<td>Record of formal schooling</td>
</tr>
<tr>
<td>Categoryid</td>
<td>Item no.</td>
</tr>
<tr>
<td>Logtimes</td>
<td>The login time</td>
</tr>
<tr>
<td>Postcount</td>
<td>The total number of posts</td>
</tr>
</tbody>
</table>
The horizontal aspect refers to the course as the researched object, aiming at describing the learning condition of all learners. For example, it describes the course access statistics (the numbers of visitors, learning hours per capita, access of each part of course, etc.). The horizontal aspect helps teachers to understand whether the course can be learned by learners and, the teacher can adjust the course content by reference to this condition.

The following is the access log on “database principles”.

This statistical log shows who is accessing the course at any time and which parts of this course is being studied. The online learning condition and attention degree of each part can be analyzed from this log to gain a comprehensive understanding of the condition and better arrange the course. From this log, the study record of each learner can be analyzed, which is conducive to understanding and mastering the learning condition and progress, and then specific guidance can be made.

4.4 Design of Association Rule Mining Module

This module mines the potential association between relative attributes of basic information such as types, gender, academic, log-in time and sum total of amount by using the Apriori algorithm. The aforementioned tables are used to form an analytical table of basic information about the learner.

| Table 3 List of candidates
<table>
<thead>
<tr>
<th>Transaction identifier</th>
<th>Project name</th>
<th>Project value</th>
</tr>
</thead>
<tbody>
<tr>
<td>00012</td>
<td>Xueli</td>
<td>A</td>
</tr>
<tr>
<td>00012</td>
<td>Xingbie</td>
<td>Man</td>
</tr>
<tr>
<td>00012</td>
<td>Studenttype</td>
<td>Elective course</td>
</tr>
<tr>
<td>00012</td>
<td>Postcount</td>
<td>a</td>
</tr>
<tr>
<td>00023</td>
<td>Xueli</td>
<td>c</td>
</tr>
<tr>
<td>00026</td>
<td>Xingbie</td>
<td>Man</td>
</tr>
<tr>
<td>00035</td>
<td>Studenttype</td>
<td>Elective course</td>
</tr>
<tr>
<td>00038</td>
<td>Xueli</td>
<td>A</td>
</tr>
<tr>
<td>00012</td>
<td>Logtime</td>
<td>8</td>
</tr>
<tr>
<td>00045</td>
<td>Postcount</td>
<td>c</td>
</tr>
<tr>
<td>00012</td>
<td>Xueli</td>
<td>B</td>
</tr>
</tbody>
</table>

| Table 4 Learners' basic information 1 - frequency item set
<table>
<thead>
<tr>
<th>Field name</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project type</td>
<td>Varchar</td>
</tr>
<tr>
<td>Value</td>
<td>Varchar</td>
</tr>
<tr>
<td>Number</td>
<td>Numeric</td>
</tr>
</tbody>
</table>

| Table 5 Learners' basic information about item 2 - frequency set
<table>
<thead>
<tr>
<th>Field name</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project name 1</td>
<td>Varchar</td>
</tr>
<tr>
<td>Value 1</td>
<td>Varchar</td>
</tr>
<tr>
<td>Project name 1</td>
<td>Varchar</td>
</tr>
<tr>
<td>Value 1</td>
<td>Varchar</td>
</tr>
<tr>
<td>Number</td>
<td>numeric</td>
</tr>
</tbody>
</table>

Data cleaning: there are data gaps and noise problems in the original data, so those problems will be cleaned.
Data generalization: discrimination of continuous data.

Data conversion: if the data too continuous, it will be difficult to find completely the same data. The table in the database is not suitable to be mined and must be converted. Therefore, the five tables will be integrated into one basic information analytical table. The data of the basic information analytical table will be insert into the “Table of information mining of learner” after discrimination and conversion.

Academic discrimination: divided into graduate student, undergraduate and college degree which are presented in A, B and C.

Log-in time discrimination: used to analyze the online learning time of the learner. It is divided into four levels: 0,8,12,20. They are respectively 0-8:00; 8:00-12:00; 12:00-20:00; 20:00-0:00.

The amount of posts is divided into four classes: a,b,c,d. They are respectively Postcount<30 pieces; 30<Postcount<60; 60<Postcount<90; 90<Postcount.

Two learner types: elective and compulsory.

The design of mining base:

Table of information mining of learner: used to store converted information of learners.

Start the data mining.

![Course access log information statistics](image)

**Figure 5.** Course access log information statistics

In this process, the following tables can be used: learning information mining table, k-item candidate set table, k-item frequent set table.

The candidate set table: used to store candidate set of each theme in the process of mining.

1-item frequent set table of basic information about the learner: used to store frequent item set of basic information about learner.
There is also a 3-item frequent set table of basic information about the learner.

Start data mining.

(1) set threshold, minimum support and minimum confidence

(2) mine frequent item set

In the candidate set, the number of each item set is counted. It is then inserted into the frequent item set table, while at the same time, the non-frequent item of the candidate set is deleted to generate the next candidate set. Taking generating frequent 1-item set as an example, the related pseudo-code is as follows: Q_L1 presents the 1-item frequent set table, Q_C1 presents candidate set table. ITEM presents the name of the item, VALUE presents the value of the item.

```sql
INSERT INTO Q_L1  //insert frequent set table
SELECT ITEM, VALUE, COUNT(*) FROM Q_C1
GROUP BY ITEM, VALUE
HAVING COUNT(*) >= MINSUPPORT;
DELETE FROM Q_C1   //delete non-frequent set from candidate set
WHERE (ITEM, VALUE) NOT IN
(SELECT ITEM, VALUE FROM Q_L1)
```

(3) After completing the operations above, the non-frequent set has been deleted from the 1-item candidate set and the rest is frequent set so that the 1-item candidate set can be connected to generate the 2-item candidate set.

(4) Steps (2) and (3) continue to be conducted until all frequent sets are found. The association rule is generated by the frequent set. Taking the learner of “network design” as an example, because there is a lack of data of some students, 560 records are selected. The threshold is set at support=0.2, confidence=0.8. Then, the association rule is observed. The rule is as shown in Figure 4.

The rules obtained from 2-item frequent set table are as follows: gender=male=online time=20(0.635,0.81), showing that 81% of male learners learn online after 20:00. The number of post=a student type II elective type (0.35,0.6) presents the 60% of learners whose sum of posting is less than 30 pieces is elective.

Through obtaining the association rule, the potential relationship between basic information and the learning condition of the learner can be analyzed to help teachers to better understand and guide learners.

<table>
<thead>
<tr>
<th>Course name</th>
<th>Course notes learning time</th>
<th>Online question answering time</th>
<th>Learning time</th>
<th>The online test</th>
<th>BBS posting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic computer</td>
<td>24</td>
<td>30</td>
<td>120</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>An introduction to aesthetics</td>
<td>13</td>
<td>0</td>
<td>630</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>Film appreciation</td>
<td>30</td>
<td>120</td>
<td>820</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>The electronic commerce</td>
<td>5</td>
<td>40</td>
<td>380</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td>The database</td>
<td>10</td>
<td>55</td>
<td>450</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

*Figure 6. Students’ study statistics*
5. CONCLUSION

This paper introduces the application of data mining for assisting teachers’ decision-making through analysis of information about learners. The problem of web-based education information being discrete and difficult to be unified, this paper firstly presents the statistical analysis module to be used to grasp the activities of learners, which helps to grasp the overall learning circumstance of the learner and understand the learning condition of this course. Then, the application of association rules for analysis of basic information about learners is described in detail. The potential relationship between basic information and activities of learners is found through association rule mining to help teachers to grasp the learning condition of learners and generate proper formative assessment.

6. REFERENCES


Susan D.J., Gary G. (1995). Distance learning and education as place, Telemetric and Informatics, 12, 231-236.

Tu J.H., Quan Z. (2014). Design and implementation the crying voice detection circuit in the baby’s supervision system, Review of Computer Engineering Studies, 1(1), 13-16.
