Association Rules Mining from the Educational Data of ESOG Web-Based Application

Stefanos Ougiaroglou[1] and Giorgos Paschalis[2]

[1] Dept. of Applied Informatics, University of Macedonia, Thessaloniki Greece
[2] Human-Computer Interaction Group, University of Patras, Patra, Greece
stoug@uom.gr, gpasxali@upatras.gr

Abstract. Many researchers have focused on the mining of educational data stored in databases of educational software and Learning Management Systems. The goal is the knowledge discovery that can help educators to support their students by managing effectively educational units, redesigning student's activities and finally improving the learning outcome. A basic data mining technique concerns the discovery of hidden associations that exist in data stored in educational software Databases. In this paper, we present the KDD process which includes the application of the Apriori algorithm for the association rules mining from the educational data of ESOG Web-based application.

Keywords: Association Rules, Apriori Algorithm, ESOG, Educational Data.

1 Introduction

Data Mining (DM) or Knowledge Discovery in Databases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections [2,3,8]. The results of the whole process are taken into consideration by experts in important future decisions in various fields.

Educational Data Mining (EDM) is an emerging interdisciplinary research area [13]. EDM deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyze educational data in order to study educational questions. The results can be used not only to learn the model for the learning process [9] or student modeling [10] but also to evaluate and to improve e-learning systems [11] by discovering useful learning information from learning portfolios [12].

Educational Data are data coming from two types of educational systems: Traditional classroom and Distance (Web-based) Education. Today, there are a lot of terms used to refer to web-based education such as e-learning, e-training, online instruction, web-based learning, web-based training, web-based instruction, etc. And there are different types of web-based systems: synchronous and asynchronous, collaborative and non-collaborative, closed corpus and open corpus, etc.

According to [13], educators and academics are responsible in charge of designing, planning, building and maintaining the educational systems. Students use and interact with them. Starting from all the available information about courses, students, usage

L. Iliadis et al. (Eds.): AIAI 2012 Workshops, IFIP AICT 382, pp. 105–114, 2012.

and interaction, different data mining techniques can be applied in order to discover useful knowledge that helps to improve the e-learning process. The discovered knowledge can be used not only by providers (educators) but also by own users (students).

KDD process, as presented in [14, 2, 3], is the process of using DM methods to extract knowledge according to the specification of measures and thresholds, using a database along with the appropriate preprocessing, and data transformation of the database. There are considered five KDD stages executed in the following order:

- Selection: This stage consists of creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed;
- Pre-processing: It consists of the target data cleaning and pre processing in order to obtain consistent data;
- Transformation: It includes the transformation of the data using dimensionality reduction or other transformation methods;
- Data Mining: It includes procedures which search for patterns of interest in a particular representational form, depending on the DM objective;
- Interpretation/Evaluation: It includes the interpretation and evaluation of the mined patterns.

The most difficult and time-consuming of the five mentioned stages are the stages of the data pre-processing and transformation. It is worth mentioning that many researchers consider these two stages as one under the label “preprocessing”.

One of the most useful and well studied mining methods is the Association Rule Mining (ARM) [5, 2, 3]. During the last decades, many scientists and practitioners of various fields apply data mining algorithms to explore and discover relations between attributes of large databases to create corresponding rules. Such rules associate one or more attributes of a dataset with another attribute between sets of items in large databases, producing an if–then statement concerning attribute values. The application of such techniques, usually, is not a simple process, but requires the successive application of the five steps of the KDD process.

ARM has been successfully applied in educational content. An interesting review of relevant applications on data of learning management system can be found in [19]. The work in [20] presents the application of ARM in the data of an e-Examination system. Other interesting relevant works can be found in [17, 18].

This work is also focused in the application of ARM in educational data. The motivation is the discovery of hidden associations that probably exist in the educational data of the database of the ESOG web-based application [1]. These data were collected by the use of the particular application from Greek secondary education students of various parts of Greece. The contribution of this paper is the application of the KDD process on the educational data of ESOG and the discovery of the corresponding association rules. For the mining process, the well-known Apriori algorithm [4] is applied through the machine learning software WEKA [6].

The rest of this work is organized as follows: The next section outlines the ESOG application as well as the data that it manages. Then, the paper presents briefly issues

that concern the association rules mining and the Apriori algorithm. Section 4 considers in details the KDD process for the export of the association rules from the ESOG data using the Apriori algorithm. The paper concludes by summarizing the whole effort and defining some certain directions for future work.

2 The Web-Based Application ESOG

2.1 ESOG System

ESOG[1] [1] is a web-based application that supports students in their school occupational guidance orientation. Particularly, students in Greece, in their last year of studies in Secondary Education (SE) have to choose the university department where they are going to continue their studies. So, according to their degrees in the annual Pan-Hellenic exams, they have to complete a form stating their preferences in University Departments (UDs). However, the knowledge domain of the UD that they choose may be out of their interest. The factors that may lead students to a wrong choice include (i) Lack of accurate information and proper guidance, (ii) Influence from relatives, (iii) selecting of a UD based exclusively on placement that it provides. Consequently, many students may face various learning difficulties during their studies. These students maybe would have a successful evolution in another science.

Thus, there is an obvious need to help students to take a right decision. ESOG is a software tool that may be useful in such cases. In particular, it is an intelligent decision support system which allows the students to indicate their degree of interest for courses have been taught in Secondary Education (SE). Then, it executes the multi-criteria analysis method ELECTRE I [7, 16] and produce a ranked based on students interests UD list.

In general, multi-criteria analysis methods [15], such as ELECTRE I, are assessments and evaluations methods based in the principle that complicated problems of decision-making, are not possible to be solved mono dimensionally. All criteria that influence the decision-making are necessary to be taken into consideration for the problem solution. The use of such methods is necessary when various parameters of a problem must be taken into account to make a decision.

The family of ELECTRE methods [7, 16] includes multi-criteria analysis methods that act into two stages. The first stage develops one or several outranking relations, which aims at comparing in a comprehensive way each pair of actions. In the second stage, an exploitation procedure elaborates the recommendations obtained in the first stage making consistent exploration and analysis in support of decision makers.

ESOG adopts the ELECTRE I method in order to aid students to decide which UD is better to choose for their studies. Specifically, the users of ESOG (students) state their liking degree (1-5) in SE courses through a questionnaire. Then, ELECTRE I is executed and the system produces the ranked list of UD groups in accordance to student interests. This list includes in the following order the ranking position of UD

[1] `http://users.sch.gr/stoug/esog`

group, the title of UD group and the total value of ELECTRE degrees that corresponds to each one of them. These values show how much each student suits with the corresponding UD group.

Technically, ESOG system has been implemented using PHP and Java Script for system intelligence and the MySQL server for the system database. It is hosted by a web server of the Pan-Hellenic School Network.

2.2 ESOG Database

ESOG Application uses MySQL for the management of its database. These data will be used by the KDD process. Figure 1 illustrates the ER diagram of the ESOG databases schema and Table 1 depicts its normalized form and lists the number of records of each table. We can observe that the questionnaire have been answered from 511 students from 14 Lyceums of Greece. Each student, through the questionnaire, expressed their interest for 24 courses (maximum for 24 courses – they may not express their interest for certain courses leaving unanswered the corresponding questions). Thus, there were collected the 10665 records of table "LIKE". Then, ESOG, by taking into consideration the 288 records of table "RELATED_WITH_1", which contains the gravity ELECTRE factors (or weights) of the 24 courses for the 12 UD groups and applying the ELECTRE I method, created for each student, 12 records in the table “RELATED_WITH_2”. The record’s field "Relation Degree" defines how much the UD groups suit in the particular interests of each student. Because of a technical problem in the database, some records of the particular table have been lost and thus it does not include 511 * 12 = 6132 registrations that should normally contain but 4839. However, this problem, as we will see in the next section, does not affect the data mining process.

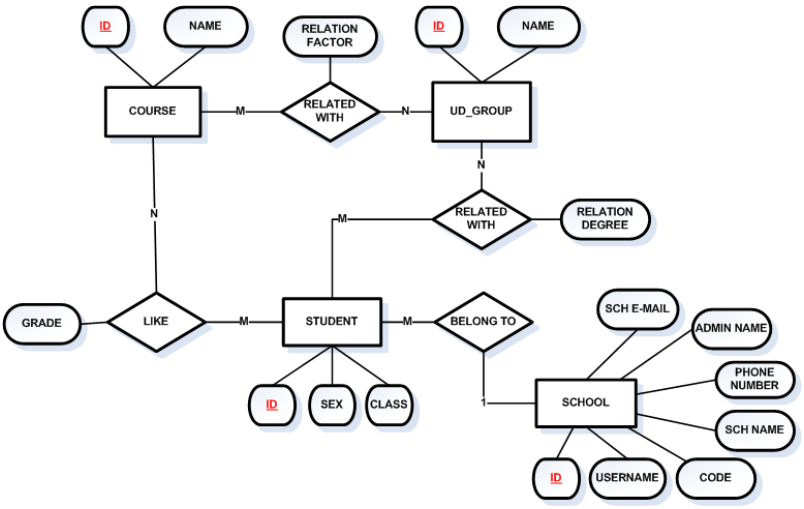

Fig. 1. ER diagram of the ESOG database

Table 1. Relational schema of the ESOG database

Table	Field	Records
COURSE	ID, NAME	24
UD_GROUP	ID, NAME	12
STUDENT	ID, Sex, Class, SCOOL_ID	511
SCHOOL	ID, E-Mail, admin_name, tel, sch_name, code, username	14
RELATED_WITH_1	COURSE_ID, UD_GROUP_ID, Relation Factor	288
LIKE	STUDENT_ID ,COURSE_ID, Grade (1-5)	10665
RELATED WITH_2	STUDENT_ID, UD_GROUP_ID, Relation Degree	4839

3 Association Rules Mining

Given a set of transactions, the problem of ARM is to discover all hidden associations that satisfy some user-predefined criteria. The association rules algorithms [5,2] solve this problem dividing the problem into two parts: mining for frequent itemsets and rules discovery from the frequent itemsets. A Frequent itemset is a set of items which are more than a threshold. The procedure of finding frequent itemsets is simple but very time consuming because of the large number of the possible combinations. The association rules algorithms differ to each other in the way that discovers frequent itemsets. Once they have been discovered, the rules production is a simple process.

A widely-used algorithm for the association rules mining is the Apripri algorithm [4]. Apriori is based on the following rule: All sub-itemsets of a frequent itemset must also be frequent. By using this rule, Apriori is able to prune a huge amount of itemsets examinations since it is certain that they are not frequent. Particularly, Apriori algorithm uses a "bottom up" approach. Frequent sub-itemsets are extended one item at a time (candidate generation), and groups of candidates are examined. It terminates when no further extensions are found. With other words, Apriori generates candidate itemsets of length l from itemsets of length $l-1$ and then it prunes the candidates which have a non frequent sub itemset. Thus, it keeps only the frequent itemsets among the candidates.

Apriori algorithm and many other data mining algorithms have been implemented under the WEKA framework [6]. WEKA (Waikato Environment for Knowledge Analysis) is popular data mining software developed in Java and distributed in a free/open source form. In the present work, we use the Apriori implementation provided by WEKA in order to mine rules from the educational data of ESOG.

4 KDD Stages for Association Rules Mining the ESOG Data

In this section, we present the five stages of KDD process on ESOG database. As we mentioned in Introduction section, we consider stages two and three as one single stage, which is described analytically in Section 4.2

4.1 Data Selection

The first stage concentrates the data used during KDD process. Therefore it defines the type of information that will be retrieved. Studying Figure 1 and Table 1, we are led to the following: We are interested in discovering association rules between the courses. Particularly, we want to examine the students' interest for a course in relation with the interest they demonstrate for other courses. Furthermore, we wanted to find if the percentage of students interested in some courses depends on the school they attend (geographical position of the school), their class and gender.

Taking into consideration the previous mentioned goals, we conclude that we have to keep the tables STUDENT, SCHOOL, COURSE and LIKE. The rest of the tables do not need to be taken into account for the KDD process, because they contain data that do not present discovery interest. It is also interesting to discover if there are relations between the UD groups and the student's answers. However, it can be reconsidered by taking into account that the student's classifications to UD groups were not directly defined by the students, but by the ELECTRE algorithm and the relation factors (weights) between the courses of UD groups.

Concerning the SCHOOL table, we have kept only the field that contains the name of school. We have done the same for the table COURSE. The LIKE table contains the student's answers in the question for each course ("How much interesting you consider course X", where X is the name of the course). The possible answers were: "very interesting", "somewhat interesting", "neutral", "not very interesting", "not at all interesting" which corresponds to a number of the likert scale: 1 (not interesting at all) - 5 (very interesting) The main problem that should be faced by the second stage of KDD process is that LIKE table constitutes an intermediate table of a many – to – many relation between the tables COURSE and STUDENT. This means that it includes many records (answers for each course) for each student.

4.2 Data Pre-processing and Transformation

In these stages, we created a file that contains the data where the mining algorithm was applied. These stages are the most difficult and time consuming. This file contains one line for each student (i.e. total 511 lines (records)), each one has 27 attributes (fields). More analytically, each file line has 24 attributes related with the student's answers in the ESOG course' s questionnaire (a field for each course). If a student did not answer a question, then the corresponding field takes the value "?". The other three fields of each line correspond to the student's sex, school and class.

The construction of the aforementioned file requires: (i) The execution of some SQL queries in the PhpMyAdmin environment provided by the Panhellenic School Network. These SQL queries retrieve the data from the corresponding database tables. (ii) The development of a small-application (written in C language) that joins the results of the SQL queries.

More analytically, an SQL query retrieved the names of courses from the COURSE table. Then they were stored in a text file. Then the contents of table LIKE were retrieved by the following SQL Query:

```
SELECT M.ID , M.class, M.sex S.name, A.COURSE_ID, grade
FROM SCHOOL S, STUDENT M, LIKE A
WHERE A.STUDENT_ID = M.ID
AND M.SCHOOL_ID = S.ID
ORDER BY A.STUDENT_ID
```

The results of SQL query were also stored in a text file. The line's number of this file is equal to the record's number of the LIKE table (for each student, there are more than one records, since they gives more than one answers – one for each course). The file records have the form of following example: <228>, <C>, <Girl>, <1st Lyceum of Nafpaktos>, <4>, <5>. The first attribute corresponds with student's code (values 1-511), while the fifth one corresponds with course's code (values 1-24). The whole line declares that student with code 228 is a girl, studying in C class of 1th Lyceum of Nafpaktos, and she is very interested (last attribute, 5: very interesting) in the course with code 4 (it corresponds with the biology course).

Finally, another file was constructed, which includes 511 records with two fields each. The first field is student's code and the second one is the number of student's answers in the questions about the courses. We remind that the questionnaire is constituted of 24 questions (one for each course) and it is not necessary all of them to be answered. The lines of this text file were retrieved by a simple group_by SQL query on table STUDENT.

Once the three text files had been created, we were able to develop a C language application for the merge of these files. In particular, this application combined the information included in three separate files and created a single final file. This new final file had 511 records and 27 attributes as well as it constituted the input of the next stage of the KDD process. Figure 2 presents algorithmically the functionalism of this small application. Considering lines 6-18, it is obvious that for each student, their answers to each course are retrieved (very interested, somewhat interested, etc). If the student did not express interest for a course, then the answer's table cell that corresponds to this course remains with its initial value (i.e. "?").

4.3 Association Rules Mining

For the association rules mining, we chose the Apriori algorithm because it is the most known and widely-used association rules mining algorithm. Furthermore, Apriori algorithm has been implemented under the WEKA framework. This fact made this stage quite simple.

Since we used an already existed Apriori implementation, this stage was quite simple. The only thing we should do was to open the file created by the previous stage in WEKA and to run the existing algorithm it provides. The goal of the current KDD stage is the discovery of association rules like the following:

> If Mathematics is very interesting and Latin course is not very interesting then Information Technology is very interesting.

```
Input:
 Text file COURSE                    ! It includes the 24 course names
 Text file ANSWERS                   ! It includes the students answers (SQL Query 1)
 Text file GROUP_BY                  ! It includes the number of answers of each student (SQL Query 2)
Output:
 Text file OUT                       ! It is the file which will be used for the mining
begin
1. Read the data of the text file COURSE and store them into the array Course[24]
2. Read the data of the text file GROUP_BY and store them into the array GroupBy[511][2]
! First line of OUT includes the names of each attributes
3. Write 'class, sex, school' to OUT
4. For each course i (i≤24)
5.     Write the course name Course[i] to OUT
6.     For each student i (i≤511)
7.         Read from txt file ANSWERS the attributes class, sex, school for the student i
8.         For each Course j (j≤24)
9.             Answer[j]←?              !Initially, the student i does not answer to any question
10.        End for
10.        For each answer τ of student i (i≤ GroupBy[i][2])   !the cell GroupBy[i][2] holds the number of answers
11.            Read the code (1≤code≤24) of the course for the answer t
12.            Read the grade (1-5) that the student i indicates for the course defined by the code for the answer τ
13.            Answer[code]←grade       !For the non answered questions the symbol ? remains
14.        End for
!Store the record of student i to OUT
15.        Write the attributes of student i: class, sex, school to OUT
16.        For each Course j (j≤24)
17.            Write Answer[j] to OUT
18         End for.
19.    End for
end
```

Fig. 2. Algorithm for the creation of the one single file which is the input for the mining algorithm

For the production of such rules, the user should define the confidence threshold that the Apriori algorithm uses through the WEKA interface. We defined the confidence threshold to be c=0.7. It means that the association rules that would be discovered would have confidence above 70%. The result of this stage is the discovery of 127 association rules[2]

4.4 Evaluation

The last stage of KDD process includes the interpretation of the rules meaning and their utilization during the educational decision-making. They can be used for the classification of students into groups according to their interests, for the investigation of interests of boys and girls as well as for ESOG questionnaire re-designing.

We have categorized the association rules mined into three groups according to their confidence[2]. The three groups are (i) strong, (ii) medium, and (iii) weak having (i) 90%-100%, (ii) 80%-89%, and (iii) 70%-79% confidence value respectively. The "strong" group includes only 4 rules. Correspondingly, the groups "medium" and "weak" include 37 and 86 respectively. Although the categorization provides the significance level of each rule, it does not mean that a strong rule is more interesting than a weak rule.

[2] List of rules available at: `http://users.sch.gr/stoug/rules.pdf`

We should mention that some of the mined rules do not present any interest. On the other hand, some of them are very interesting and not expected. Some of the most interesting rules and conclusions are:

- The majority of association rules indicated that students who dislike one or more courses also dislike another one course. So, many rules have “1: not at all interesting” as answer in the corresponding question.
- There were three courses (Foreign Languages, Aesthetic course, Physical Education) that that did not take part in any rule. It means that they do not relate to any course or other attribute, i.e. gender, school and class.
- At the most schools, students can be classified into two categories: (i) sciences, and (ii) humanities. This observation ensures us about the student tendency to be divided into the particular categories.
- Boys interests participated only in 4 rules, while girls interests participates in 25. It may mean that, contrary to boys, girls have similar interests.
- Some rules indicate that girls are less interested in sciences courses than the humanities (theoretical) ones. On the other hand, boys are interested in technology and usually dislike the theoretical courses.
- The rule “if programming is very interesting then Informatics is very interesting” indicates to the questionnaire designers that these questions should be merged since the students consider these courses as similar.
- The Technology course is not presented in any rule with that of Informatics or Programming. It may mean that students consider the Technology field as different with that of informatics.
- The student’s tendency to declare their "dislike" for the courses "Latin" and "Agronomy" answering "not at all interested" in the corresponding questions of questionnaire, constituted the main factor for many rules construction for these particular courses.

5 Conclusion

In this paper we presented the KDD stages for the association rules mining the ESOG database which contains educational data. This process produced 127 association rules that could help and guide Greek Educators and School Managers to make educational decisions, design learning activities according their student’s interests and efficiently manage the classroom (divide class into groups of students with similar interests, adapt course’s content etc).

During the conduction of this work, many questions arose that indicated directions for future research. One of these directions is the application of other types of data mining algorithms in the ESOG database (classification or clustering algorithms).

References

[1] Ougiaroglou, S., Kazanidis, I.: Occupational Guidance through ELECTRE Method. In: Leung, H., Popescu, E., Cao, Y., Lau, R.W.H., Nejdl, W. (eds.) ICWL 2011. LNCS, vol. 7048, pp. 142–147. Springer, Heidelberg (2011)
[2] Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science (2011)
[3] Margaret, D.: Data Mining: Introductory and Advanced Topics. Prentice Hall (2003)
[4] Agrawal, R., Ramakrishnan, S.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, pp. 487–499 (1994)
[5] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C. (May 1993)
[6] Ian, W., Eib, F.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
[7] Figueira, J., Mousseau, V., Roy, B.: ELECTRE methods. In: Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 133–162. Springer (2005)
[8] Klosgen, W., Zytkow, J.: Handbook of data mining and knowledge discovery. Oxford University Press, New York (2002)
[9] Hamalainen, W., Suhonen, J., Sutinen, E., Toivonen, H.: Data mining in personalizing distance education courses. In: World Conference on Open Learning and Distance Education, Hong Kong (2004)
[10] Tang, T., McCalla, G.: Student modeling for a web-based learning environment: A data mining approach. In: Eighteenth National Conference on Artificial Intelligence, Menlo Park, CA, USA, pp. 967–968 (2002)
[11] Zaïane, O., Luo, J.: Web usage mining for a better web-based learning environment. In: Proceedings of the Conference on Advanced Technology for Education, Banff, Alberta, pp. 60–64 (2001)
[12] Hwang, W., Chang, C., Chen, G.: The relationship of learning traits, motivation and performance-learning response dynamics. Computers & Education Journal 42(3), 267–287 (2004)
[13] Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications 33(1), 135–146 (2007)
[14] Fayyad, U.M.: Data mining and knowledge discovery: making sense out of data. IEEE Expert 11(5), 20–25 (1996)
[15] Roy, B.: Classement at Choix en Presence de Points de Vue Multiples (la Methode ELECTRE). Rev. Franc. Inform, et Rech. Oper. 2(8), 57–75 (1968)
[16] Figueira, J., Greco, S., Ehrgott, M.: Multiple Criteria Decision Analysis: State of the Art Surveys. The International Series in Operations Research and Management Science. Springer (2005)
[17] Merceron, A., Yacef, K.: Interestingness measures for association rules in educational data. In: Proceedings of Educational Data Mining Conference, pp. 57–66 (2008)
[18] Romero, C., Romero, J.R., Luna, J.M., Ventura, S.: Mining Rare Association Rules from e-Learning Data. In: Proc. of Educational Data Mining Conference, pp. 171–480 (2010)
[19] García, E., Romero, C., Ventura, S., Calders, T.: Drawbacks and solutions of applying association rule mining in learning management systems. In: International Workshop on Applying Data Mining in e-Learning (ADML 2007)), Crete, Greece, pp. 15–25 (2007)
[20] Mamcenko, J., Sileikiene, I., Lieponiene, J., Kulvietiene, R.: Analysis of E-Exam Data Using Data Mining Techniques. In: Proc. of 17th International Conference on Information and Software Technologies (IT 2011), Kaunas, Lithuania, pp. 215–219 (2011)