Discovering association rules of information dissemination about Geoinformatics university study

Zdena Dobesova

Department of Geoinformatics, Faculty of Science, Palacky University,
17. listopadu 50, 779 00 Olomouc, Czech Republic
zdena.dobesova@upol.cz

Abstract. The article presents the data mining of dataset about spreading information among the university study applicants. The data were collected during admission procedure of applicants for bachelor study branch Geoinformatics and Geography at Palacky University in Olomouc (Czech Republic). Answers were received by questionnaire in two years, 2016 and 2017. Data collecting and processing aimed to discover the dissemination of first information about this specific specialization among graduates at secondary schools. Statistics and data mining techniques, namely finding association rule were used. Data mining discovered some unexpected relation and association in the data. Interesting results bring feedback about the impact of various presentation activities like Open Day, GIS Day or publishing information on the Internet. Results will also be reflected in future advertisement strategies of the study branch Geoinformatics to assure increasing interest of the potential applicants.

Keywords: Data Mining, Associations Rules, Questionnaire, Geoinformatics, Classification, Advertisement, University Applicant.

1 Introduction

The right choice for future profession and specialization is a fundamental decision in the life of young people. This decision is mainly connected with finishing common secondary school and continuing study at university. The information dissemination about interesting study branches is important both for applicants both for universities. The importance for the universities is the fact that the number of potential students in the Czech Republic has been decreased due to demography evolution. Recently, some study branches have not filled up the declared quota of students. The best result is when personal interest of student corresponds with the offer of study to assure his professional success in career.

The Department of Geoinformatics has been collecting data about spreading information to applicants. The reason was to explore the penetration and influence of advertisement and other activities to the potential university students of geoinformatics. The generation of association rules, one of data mining technique, was used [1, 2]. The advertisement activities and collection of data are described in section 2. The processing data by association rules is described in section 3.

2 Information about university study

Delivering information about offered study branches at universities has several various ways. Some of them are common, and others are specific for the Czech Republic or Palacky University in Olomouc.

2.1 Advertisement strategies about study branches at university

Palacky University issued printed leaflets and booklets about a list of offered study branches. The materials are delivered to the secondary schools, especially to the teachers and students. Detail study information is also presented in digital format on web pages of universities to describe the structure and aim of study braches. Description about specializations is supplemented by information about admission procedure, deadlines for application, fees and opportunities to forgive enrol exams.

Sometimes, the students actively try to find interesting study branch and information about the structure of the study. They ask their teachers at secondary schools for information about study opportunities at universities. Students also debate with their parents and siblings about the future profession. The recommendation and experiences of friends and schoolmates are very influential for their final decision. Students from the Czech Republic actively attend special study fair Gaudeamus that is organized in Brno and Prague in the fall and winter of each year [3]. Moreover, the Gaudeamus fair takes place in Nitra and Bratislava in the Slovak Republic. All national and abroad universities present their various study programs at these fairs.

Moreover, universities organize their form of public presentation of studies. Palacky University organizes annually two Open Days, one in November (Friday) and second in January (Saturday) [4]. Applicants could attend the university to personally speak with future teachers and receive information about admission procedure. Many excursions are arranged to present departments, classrooms, laboratories, equipment and research results of scientific teams. The excursions are arranged both during Open Days and optionally during the whole academic year.

2.2 Information about study of Geoinformatics

Department of Geoinformatics at Palacky University issued a unique collection of printed leaflets about bachelor study branch Geoinformatics and Geography and master study Geoinformatics. The aim is to spread information to future potential students to the secondary schools and the public. Leaflets contain a description of the aim of study and list of subjects with highlighted topics. One leaflet “How live geoinformatics professionals in practice?“ brings interviews with geoinformatics graduates about their different jobs. Several other leaflets exist about department and scientific research. They describe research projects, published books, map and atlas production (Fig. 1).

Fig. 1. Collection of leaflets about Geoinformatics study and research

Study branch Geoinformatics is presented at all presentation days like Open Days and Gaudeamus Fairs. Moreover, the study is presented at special action named GIS Day. GIS Day is an annual worldwide action that presented geographic information systems (GIS) technology. It is regularly organized at third Wednesday in November. Schools, firms, and many organizations prepare many presentations for the public to discover and explore the benefits of GIS. GIS Day a good initiative for people to learn about geography and the uses of GIS [5]. More than one thousand participants take part at this event around the whole world. The Department of Geoinformatics regularly organizes a presentation for a student from secondary schools from Olomouc and other neighbouring cities. Groups of students with their teachers of geography attend computer laboratories at the department and try to use GIS software, prepare maps or solve a geographical task. They make an idea about utilization of GIS technology, and they receive basic information about opportunities to study Geoinformatics discipline. Present students of master study Geoinformatics also attend secondary schools with presentation lectures about study opportunities. These activities run during the whole school year in various cities and schools.

Besides these all presentations activities, other publicity channels exist. The Teacher News (www.ucitelskenoviny.cz) is issued in the Czech Republic. The autumn issue regularly brings a complete list of Czech Universities and all study branches offered for next academic year. Next information channel is radio and TV. Local and national radio and television broadcasting (like Český rozhlas, Radiožurnál, ČT1, ČT24) bring several reportages about Palacky University and interviews with

famous researchers and teachers. These entire information channels could be a source of first information or a hint about particular study branch.

It is very difficult to evaluate the influence of all mentioned activities to the attraction of potential applicants. The research task was to find the impact of presentation days, leaflets and other channels of advertisement to the spreading of information to the future applicant of study branch Geoinformatics.

3 Data acquisition and data mining

The data have been collected from applicants for evaluation of the advertisement impact. The Department of Geoinformatics prepared a questionnaire for applicants to imagine the way of spreading the first information about the study. The statistical methods and data mining methods were used for processing data from the questionnaire. Data mining is the process of discovering and extracting hidden patterns from different data types to guide decision makers and making decisions [6]. Data mining performs different methods including classifications, clustering, regression, association rule discovery, decision tree, and pattern recognition. The method of association rules generation was chosen because the received transactional data had a suitable form for this mining method. The rules could be stored and reused as knowledge in intelligent and expert systems [7]. Clustering and classification of data, as frequent data mining methods, was used as next step of data analysis [8].

3.1 Questionnaire and data matrix creation

The data were collected during admission periods in two years 2016 and 2017. The data was gathered by questionnaire from April to September each year. It is hard to meet all applicants at the same time. Some applicants only send the application in February without starting the study and any other connections to the university.

Some attended applicants were first asked to fill the questionnaire in April on information weekend named “GIS weekend” organized by Department of Geoinformatics. This meeting with applicants was an introductory weekend about enrolment tests, information about the structure of the study. The presence was optional for applicants, so the number of received questionnaires was approximately 50%. Subsequently, the questionnaire was delivered during enrol exams in June to the remains students. Finally, the last group of the student was questioned in September (about 8%) at the beginning of the study. These students had forgiven enrol exams, and they did not attend GIS weekend in April. The groups of students were punctually identified to prevent duplicated in filled questionnaires. This way of the questionnaire spreading covered nearly all applicants except them that they did not attend exams and did not start the study of Geoinformatics (about 10%). They only sent the application in enrolment period without any other interests.

The structure of questionnaire was straightforward. The necessary information like name of the student, secondary school, and the city was filled with the introduction section. The main question was:

„How did you find out information about studying Geoinformatics at Palacky University?"

There were 12 possible answers:

- Teacher of Geography at secondary school *(TeacherG)*
- Teacher of Information science at secondary school *(TeacherI)*
- Lecture at secondary school provided by Department of Geoinformatics *(Lecture)*
- Attending of GIS Day at Department of Geoinformatics *(GIS Day)*
- Open Day at Palacky University (November and January) *(Open Day)*
- Friends, schoolmates, siblings *(Friend)*
- Parents and grandparents *(Parent)*
- Gaudeamus Fair in Prague or Brno *(Gaudeamus)*
- Leaflets about study branch Geoinformatics *(Leaflets)*
- TV and Newspapers *(TV News)*
- Internet by your search *(Internet)*
- Teacher News *(T News)*

Each type of answer was assigned with the code. The codes of answers are in Italics in brackets above. The codes are used in the graph in Table 1, 2 and Fig. 2.

Some student`s responses contain one or more answers. The maximum was six answers by one applicant; it means six various sources of information. The most frequent was two or three sources of information about the study. The answers from the questionnaire were collected like transactions in the table. The term of the transaction is used in Market Basket Analysis (MBA) to discover association rules in data mining process [1, 2, 9]. Students were assigned by identification number ID. The example of input data for data mining is in Table 1.

Table 1. Input transactional data example.

ID	Items
1	TeacherG, Open Day, Internet
2	GIS Day, Gaudeamus, Leaflets, Internet
3	Friend, Internet
4	Parents, Gaudeamus, Leaflets
5	Lecture, Open Day, Internet

The transactional table data were converted to the matrix where items are binary attributes. Attribute value 1 represents the presence of answer; the 0 represent the absence of an answer. The resulting matrix is very sparse matrix due to the low frequency of answers (a lot of zero values). The example of the data matrix is in Tab. 2. The data in Tab. 2 correspond to the transactional data in Tab. 1. There are not all 12 possible answers (attributes) to reduce the size of the table in this article. The conversion of input data to the matrix was source data for statistics and association rules generations.

Finally, the data was supplemented by the information, if students started or did not start the study of Geoinformatics in September. It was taken from evidence about new students in the first grade. It is also interesting the association of information dissemination about geoinformatics study and the real start of the study. Moreover, information about attending the GIS weekend was added to the data form questionnaire.

Table 2. Input matrix example for association rules.

ID	Teacher G	Lec-ture	Open Day	GIS Day	Friend	Parents	Gaudea-mus	Leaf-lets	Internet
1	1	0	1	0	0	0	0	0	1
2	0	0	0	1	0	0	1	1	1
3	0	0	0	0	1	0	0	0	1
4	0	0	0	0	0	1	1	1	0
5	0	1	1	0	0	0	0	0	1

3.2 Data analyzing

Firstly the basic statistics were prepared. A total number of respondents was 46 in the year 2016 and 55 in 2017. The number of answers for each type of information for two years is displayed in Fig. 2. The highest number of answers has the source form Internet and web pages. The answers are expectable because of the primary information about bachelor study, an overview of subjects, an organization of study and examples of enrolling test (mathematics and geography) are placed on web pages Department of Geoinformatics (www.geoinformatics.upol.cz). Also, faculty web pages contain necessary information about admission procedure, conditions, and dates (www.prf.upol.cz). Also, other sources from the Internet could be assumed like Facebook. The particular web pages were not inquired in the questionnaire because the answers would not be reliable. This answer could be assumed to be biased and over evaluated by the time delay between received first information and time of survey (sometimes more than six months). Also, there is a bias about the primer information about the existence of study branch Geoinformatics on the Internet. Primer information could be outside of the Internet, but applicants subsequently looked for detailed information on the Internet. It is hard to remember for applicants after two or six months to distinguish if the information form Internet was solely primary information. However, the answers verified that the Internet is also the most frequent information source about the study. Surprisingly, a minimal number of answers got: teacher of information science subject, TV and newspapers, Teacher News. Very influential are Open Day, Gaudeamus, Lectures at secondary schools and Leaflets. Fiends and siblings have more answers than recommendations by parents.

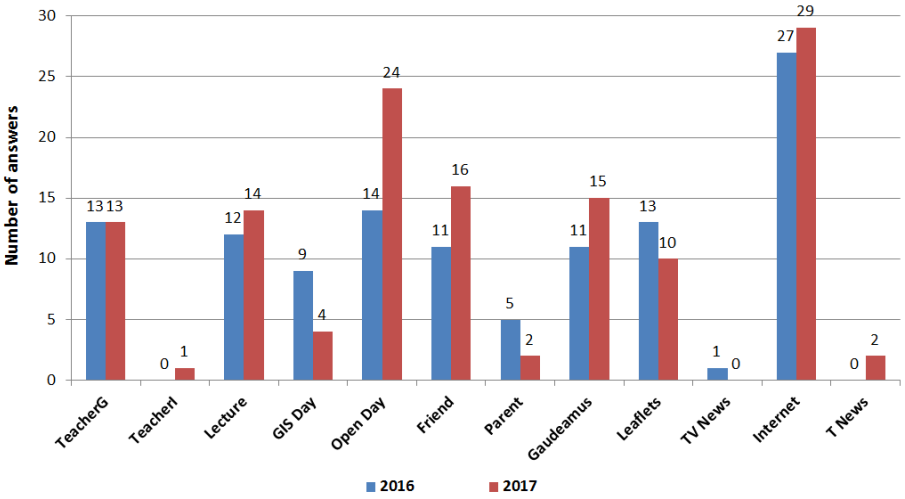

Fig. 2. Graph with numbers of responses for each source of information about the study.

3.3 Association Rules

Association rule discovery task is a typical example of unsupervised learning [6, 9] where the aim is to discover the correlation among attributes in the transaction data like presence-absence data received by questionnaire. Based on converted data to the matrix, association rules find relations among items. The association rules are displayed on the next form (1), where the antecedent is a precondition of rule and succedent is the result of the rule. The interpretation of the rule is: When the precondition is valid than also the result is valid [9]. The inference made by an association rule does not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between items in the antecedent and consequent [1].

$$Antecedent \Rightarrow Succedent \quad (1)$$

The evaluation of the validity of association rules depends mainly on three important values called support, confidence and lift. Support represents the frequency of the items in datasets, in other words, objects that fulfill both antecedent and succedent. The confidence indicates how often the rule has been found to be true in itemsets that represent this rule. Lift is the ratio of the probability when antecedent and succedent occur together to the two multiple individual probabilities of antecedent and succedent [1]. Antecedent and succedent are independent when the value of the lift is 1. The higher value of lift means that existence of antecedent and succedent is not just a random occurrence, but because of some relationship between them.

In the process of the rules generation is user-set two thresholds: minimum support and minimum confidence. The process of generating association rules consists of two steps. Firstly, to find all frequent itemsets and secondly generate the association rules

from frequent itemsets. The recommendation for minimum confidence is about 70% for data obtained by questionnaire [2], and minimal support 5%. The Apriori algorithm is the most well-known algorithm for finding the frequent itemsets and Boolean association rules [10]. The software WEKA v 3.8 was used for data mining of presented questionnaire data. The Apriori algorithm is implemented in WEKA.

Both the matrix with one-zero (presence-absence) information, both matrix only one-one (presence) information was processed. The matrix with presence-absence data produces a lot of strong rules (high support and high confidence) but only with absence items. The sparse matrix did not produce a lot of frequent positive, strong rules.

The extracted presence-absence rules are briefly described like this:

- ***Parent=0 TeachN=0*** *=>* ***TeacherI=0 TVnews=0*** (*support 92%, confidence* 99 *%*)
  This rule means that many students do not receive any information from Parents, Teacher News, and Teachers of Informatics science. All these channels are a weak source of information with high occurrence and high confidence. It is the strong rule.

The rules with lower support and some presence-absence items are like these:

- *TeacherG=0* ***Gaudeamus=1*** *TeachN=0 => Lecture=0 GIS_Day=0* ***OpenDay=1*** *Friend=0* (*support 17%, confidence 59 %, lift 2.7*)
  This rule means that visiting Gaudeamus Fair and visiting Open Day has high co-occurrence without other information. The rule covers a group of students that have no information from teachers of geography, information lectures, friends and GIS Day, so they very often attend the public presentation, and they prefer the personal receiving of information by humans. They travel to the fair and the university to find some information.
- *TeacherG=0 Lecture=0* ***Gaudeamus=1*** *TeachN=0 => TeacherI=0 GIS_Day=0* ***OpenDay=1*** *Friend=0* (*support 14%, confidence 71 %, lift 2.67*)
  This rule is near the previous one. Missing information from teachers of geography, lecture at secondary school, Teacher News and friends forces the students to attend the Gaudeamus and Open Day personally.
- *Gaudeamus=0 and Open Day=0 =>* ***Internet=1*** *(support 29%, confidence 52%, lift 1,02)*
  This rule expresses that student without any information from campaign use namely the Internet. The absence at the Gaudeamus Fair and Open Day (no personal attendance) and searching information on the Internet, on the other hand, are an independent phenomenon according to the lift.

At the second step, only presence matrix of items was used. The presence of an item in a transaction is considered more important than its absence in that case. The list of interesting rules is:

- ***Friend=1 Leaflet=1*** *=>* ***Internet=1****(support 7%, confidence 86%, lift 1.55)*
  This rule could be considered as a searched exception. It has low support and high confidence. Students that have information from friends and read the leaflets also use the Internet. These students did not attend public presentation like Fair.
- ***TeacherG=1*** *=>* ***GIS_Day=1*** *(support 26%, confidence 27%, lift 2.09)*

This rule expresses: If a student has information from Teacher of Geography at secondary school they also attend (in 27% of cases) GIS Day. According to the lift 2.09, the information from the teacher has high co-occurrence with GIS Day.

- ***Friend=1 => OpenDay=1*** *(support 27%, confidence 30%, lift 0.79)*
  This rule expresses that student with information from friends also attends Open Day. It is not a frequent rule but the preferable rule.

Finally, the generation of association rules from the data with the supplemented information about the start of study brings some interesting rules. There is no strong rule with the very high support and confidence. Discovered rules are:

- ***OpenDay=1 Internet=1*** *=>* ***Study=1*** *(support 20%, confidence 75%, lift 1.33)*
  That means that students who attend Open Day and have information on the Internet they start the study.
- ***Friend=1 and Internet=1 => Study=1*** *(support 13%, confidence 85%, lift 1.5)*
  It is the rule with the low support and high confidence in case of started study. It could be assumed as an exception that student starts study based only the recommendation of friends and the Internet.
- ***Internet=1 => Study=0*** *(support 13%, confidence 85%, lift 1.5)*
  This rule revealed situation when a student had information only from Internet he did not start the study. The Internet as a sole channel of information is very weak, and students are not interested in the study of Geoinformatics.

4 Results

The evaluation of the information dissemination about study brings some interesting information. The Apriori algorithm in WEKA software was used for the generation of association rules. Gaudeamus Fair, University Open Day and Internet are the most influenced activities to the applicants. If students attended Gaudemus Fair, they also attend Open Day and use the Internet. Personal meeting of the applicant with teachers of the university was very influential. Fair and Open Day are frequently used way of collecting information by secondary school graduates. These ways of personal dissemination of information form the first group of active applicants. The second group of applicants used the only Internet and they did not attend Gaudeamus and Open Day. The addition of the information about the start of the study revealed that students with information based only on Internet very often did not start study (support 13% and confidence 85%). High support 92% were discovered in rules with the absence of information from teachers at secondary schools, Teacher News, TV news and parents. These channels of information are very weak. Interesting

Surprisingly, the recommendation and experiences of friends and schoolmates are a very influential way of receiving information and final decision to start the study. The combination of friend, leaflets, and the Internet is interesting rule exception with small support but high confidence. This rule describes a small third group of students that prefer private searching of information. Also, high confidence has a combination of Open Day and Internet source to start the study. The Internet as a sole source of information did not assure the start of the study. Detection of this isolated source af-

firms that more channel of information is necessary to assure interest of graduates and successful start of the study.

The discovered rules help improve and intensify some forms of spreading information about the study. The most important is active and punctual delivering information during Gaudeamus Fair and Open Day about the study branch Geoinformatics. Other activities like GIS Day, lectures at secondary school are less frequent but could be an indirect source for friends, schoolmates and teachers of geography. These people potentially hand over the information to the target group of applicants. Finally, the actual and complete university web pages about the study are assumed as the necessary essential base of information.

Application association rules as one of data mining techniques is a new approach in processing questionnaire survey about dissemination information about the university study. The comparison with other study branches and universities is impossible due to lack of reliable information. This area opens new opportunities for retrieving interesting information about future applicants.

5 Acknowledgment

This article has been created with the support of the Operational Program Education for Competitiveness – European Social Fund (project CZ.1.07/2.3.00/20.0170 Ministry of Education, Youth, and Sports of the Czech Republic).

References

1. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc. (2005)
2. Šarmanová, J.: Methods of Data Analysis (Metody analýzy dat). Technical University of Ostrava, Faculty of Mining and Geology, Ostrava (2012), (In Czech)
3. MP-Soft, Gaudeamus Fair, Available at https://gaudeamus.cz/
4. Palacky University: Why study at Palacky University? (Proč studovat na Univerzitě Palackého?), Available at: www.studuj.upol.cz (In Czech)
5. Esri, GIS day, http://www.gisday.com/
6. Witten, I.H.: Data mining. Burlington, Morgan Kaufmann (2011)
7. Brus, J., Dobesova, Z., Kanok, J., Pechanec, V.: Design of intelligent system in cartography. In: 9th Roedunet International Conference (RoEduNet), 2010, pp. 112-117. (2010)
8. Petr, P., Krupka, J., Provaznikova, R.: Statistical Approach to Analysis of the Regions. In: Fujita, H., Sasaki, J. (eds.) Selected Topics in Applied Computer Science, pp. 280-285. World Scientific and Engineering Acad and Soc, Athens (2010)
9. Pavel, P.: Methods of Data Mining (Metody Data Miningu), part 2, University of Pardubice, Faculty of Economics and Administration, Pardubice (2014), (In Czech)
10. R. Agrawal, R. Srikant: Fast Algorithms for Mining Association Rules in Large Databases. In: 20th International Conference on Very Large Data Bases, 478-499, 1994.