Mining Retention Rules from Student Transcripts: A Case Study of the Information Systems programme at a Federal University

Carlos V. A. Silva[1], Marcelo S. Santos[1], Daniela B. Claro[1], Veronica M. C. Silva[1], Marcos Silva[1], Silvana Ribeiro[1],Ana R. Telles[1], Denivaldo Lopes[2]

[1]Instituto de Matemática – Universidade Federal da Bahia
(UFBA)
Av. Adhemar de Barros S/N – 40.170-110 – Salvador – BA – Brazil

[2]Departamento de Engenharia de Eletricidade – Centro de Ciências Exatas e Tecnologia
Cidade Universitária – Federal University of Maranhão (UFMA)
65080-040 – São Luís – MA – Brazil

carlos.andrade@acm.org, marceloss@dcc.ufba.br,{dclaro,cadena}@ufba.br

massilva@dcc.ufba.br,{silvanar,anaregi}@ufba.br,dlopes@dee.ufma.br

Abstract. *Due to the increase in inflows, mainly because of REUNI procedures, and low completion rate also observed in several programmes on Brazilian universities, it is necessary to identify which factors may cause the students to remain in their programmes longer than expected or even leaving the university before their conclusion. In this work, we hypothesize that the combinations of classes that the students have to do in each semester is not appropriated and can cause student retention, leading to a huge loss to the university. Thus, we present a case study of analyzing retention rules in the Information Systems programme at the Federal University of Bahia. With the resulting rule set obtained from mined student transcripts, we discuss how changes can be made in a programme so as to decrease the student retention rates.*

1. Introduction

Brazilian universities have seen an increase in the number of inflows, mainly because of REUNI procedures (Restructuring and Expansion of Federal Universities) [REUNI 2012]. In its design and implementation at the Federal University of Bahia (UFBA), it was estimated in 2012 an enrollment of over 32,000 students because of new programmes, of which more than 10,500 of them are enrolled in evening programmes. Indeed, in 2012, 29,171 students were enrolled, of which 3,313 were enrolled in BI (Interdisciplinary Bachelors) and 6,102 students were enrolled in evening programmes, of which 2,107 refer to an evening BI programme. Among the main guidelines of REUNI, it is possible to highlight: an increase in the number of places of entry, especially for evening programmes and a reducing rate of evasion and occupation of unfilled vacancies. According to [Filho et al. 2010], within the REUNI implementation, the completion rates expected was about 90% of the inflows. But completion rates in undergraduate programmes are still much lower than the desired.

Among the factors that contribute to the low completion rates, it is possible to highlight [Manhães et al. 2012]:

DOI: 10.5753/CBIE.SBIE.2013.577

1. the evasion of programmes and
2. the not completion of a programme in a regular period, thus remaining students in programmes beyond the period.

In this work we stated that a student is retained when he fails a particular class in a previous semester and he cannot do a certain class in the following semester. Failure to complete the regular period can be due to the retention caused by a particular class, damaging the regular flow of the academic student's life. The non regular flow of academic student's life cause loss to the University, especially in programmes under implementation. This happens because the university must offer the required classes per semester, thus moving teachers to a ratio far below the desired by REUNI [Filho et al. 2010] 18:1 (18 students to 1 teacher). Furthermore, when a student fails a particular class he has to repeat it. According to [Barros and Mendonça 1998], despite the positive effects of student's failure in a course, it is possible to highlight two negative effects related to it: i) it obstructs the regular educational flow, raising the cost of a student from an University's perspective. To the family's point of view, the student spends more than the average time of the programme to have a professional position and as a consequence a financial return to aid their family, ii) student's failure has a negative effect on self-esteem and motivation, increasing the likelihood of future failures. The failure creates a subclass of students where every new failure increases this selected and underperforming subgroup of students.

Thus, the present work has as main objective to analyze the academic performance of students in order to highlight which classes and which factors are responsible for the retention and that may be hindering the completion curriculum in a regular time, specifically at the Federal University of Bahia. Technically, it was aimed to mining the data available in the Academic System analyzing each student transcript and each class followed by each student. Through Data Mining techniques [Tan et al. 2009], specifically the association rules, it was possible to identify some classes that were causing student's retention. This work dealt only with the Information System programme at the Federal University of Bahia in order to first make a preliminary examination of our proposed algorithms and then validate our proposal.

This work is structured as follow: Section 2 presents related work, section 3 depicts our dataset. Section 4 discusses our results and section 5 presents the programme threats of validity. Section 6 presents our conclusion and future work.

2. Related Work

A couple of related works have already been published in attempting to better understand retention patterns using different data mining algorithms and hypothesis. Nandeshwar et al.[Nandeshwar et al. 2011] conducted a literature review and suggested using a suitable amount of different learning methods. They found that student's family related variables, socio-economic status, high GPA (Grade Point Average) and exam grades had influenced over the student retention problem. Zhang et al. [Zhang et al. 2010] assembled a data-warehouse of three systems (library, online learning and academic), and used Naive Bayes in order to notify students that there is a potential chance that he will be retained. Yu et al. [Yu et al. 2010] used decision trees and social factors to identify the impact the retention. Campello et al.[Campello and Lins 2008] analyzed the duration of 6

years student's socio-economic, enrollment exam and student transcripts. They used clustering algorithms to identify profile of students that were retained at Federal University of Pernambuco. [Manhães et al. 2011] used ten different data mining algorithms to identify early those who tend to evade the course of Engineering of the Polytechnic School of UFRJ using as information the high GPA and students' grades in the class of the first semester.

We can observe that most works focus on the hypothesis that socio-economic and student profile factors are the main cause of university retention. This differs from our hypothesis. Here, we are interested in observe if the recommended classes per semester have any impact in student retention problem, which to the best of our knowledge has not been addressed previously. Lastly, it is important to note that the most identified related works used an abroad dataset to analyze the retention problem. This highlights the need of research in mining retention problems in national universities.

3. Dataset

There are three entities in our dataset: students, classes and programmes. A student is identified by an id, a date of birth, a citizenship and nationality. A class is identified by a code (e.g. MATA02) followed by a name (e.g. Calculus A) and the amount of class hours in a given semester. A programme can be identified by a name (i.e. Information Systems) followed by a code.

A student is enrolled in a programme, which in our case study we started by the Information Systems programme. This enrollment occurs each semester. The student may also quit the programme for varying reason or be removed from it. It is also registered the reason for the student leaving a class. During enrollment in a programme, the student is assigned a recommended curriculum. He will also have a GPA, and the total class work hours he has done up to the most recent semester.

To complete a programme, a student has to do a set of classes. Some classes are required by the programme. Others classes the student can choose from a list, depending on his interests. Also, a class may vary over the recommended enrollment semester depending on the programme.

Finally, a student may also request classes on any given semester. In case of a regular semester, the student may have different remarks, such as: be approved in a class, be failed in a class due to not attending it, be failed due to grade, had the class transferred from another institution, or transferred within the same institution from a different programme. Internal class transfers are assigned a grade, while external class transfers are not. If the student is approved or is failed due to grade, his grade is assigned, otherwise no grade is assigned. Nevertheless the student can request to interrupt the semester for varying reasons. We are only able to distinguish health reasons from the remaining.

In order to analyze the dataset of those students, it is important to define which granularity we do use to tackle the retention problem.

3.1. Granularity

The finest grain of data in our dataset is *(student,semester,class)*. That is, which class a student has taken in a given semester. Analyzing the data in such grain provides insight

of which classes may be more associated to the student being retained. Specifically, this analysis emphasizes the hypothesis that *student retention can be caused by the combination of classes on each semester*. From this point of view, modifying the syllabus flow chart would diminish the chance of a student being retained (given that the student follows the recommendation).

We have analyzed two other views of granularity: *(student,semester)* and *(student,class)*. In the first granularity, our dataset had a few aggregation on *(student,semester)* grain, thus explaining how many students are doing a given class in a given semester (crowded classes), and how well a student stands on each of the classes in respect to his classmates. Other potential variables would include the student's semester workload in a given semester, number of intensive classes in a given semester, etc. Specifically, results obtained in this grain would lead to the hypothesis that *the cause of student retention is associated to the student instead of the organization of the classes on each semester*. It also emphasizes and suggests that the retention cause may occur over time.

One last possibility for grain would be *(student,class)*. The critical problem here is that we no longer know which classes the student has taken in parallel or which classes, providing little insight to discover retention patterns.

3.2. Transformations

As our choice was the most finest granularity *(student,semester,class)*, we performed three programme transformations, which we describe in this section.

First transformation. Since our grain includes classes and semesters, and association mining requires categorical variables, we do categorize the variables. Despite the nature of our variables not being categorical, we believe that such dichotomization does not bias the result. In particular, the decided intervals for the dichotomization were made based on interviews from domain experts who had experience in dealing with all programmes of our university.

Second transformation. From the semester the student enrolled in his programme, we can understand each new semester as of being his second, third, forth, semester etc, which can be understood as the student *relative semester* to his year of enrollment (e.g. 2009.1 , 2009.2, 2010.1). We use the *relative semester* for sub-setting the dataset. Notice that this is different than sub-setting by year of enrollment. Specifically, if we sub-set by semester of enrollment, students of different year of enrollment should not occur within a given sub-set, while in a given sub-set of *relative semester* can contain students of different years of enrollment, so long both students have been long enough on the programme to have in common a second, third semester, etc. The net effect in the dataset of this sub-setting is that, given that no student graduated yet in the information systems programme, the further we move from the students first semester, the smaller the dataset is (as less students have made far enough to the fifth semester, and so on).

We decided on use the *relative semester* sub-setting because it seems intuitive to test our hypothesis that changes can be done to the programme flow chart in order to reduce retention. These changes should be made in such a way that benefits students of any year of enrollment when they are on their respective *relative semester*.

Third transformation. Lastly, we should be aware that a student may be enrolled

on a given class for his second, third or fourth time (maximum amount of times that a student is allowed to fail *each* class). For instance, a student, on his second and third *relative semester* may have two different rows that he took Calculus A. It is of our interest to know which one was the student second, third and fourth attempt for taking Calculus A or any other class when creating our rule set. To do so, we transformed each of our variables so that their label indicates the attempt count (e.g. Calculus A_1, Calculus A_2 instead of Calculus A for both rows).

We initially obtained three tables (second, third and forth *relative semester*) with respectively N = 92, N = 91, N = 54 table sizes. For each row we had a student on his *relative semester*, and on its columns we had all the classes the student could potentially do up to four times. Values could be passed, failed, or failed due to attendance frequency. We also experimented using other variables such as class workload, GPA, amount of enrolled classes, but they usually led to obvious results (e.g. a student with high grade was not retained).

The final three tables used for analysis contain therefore all possible classes a student can enroll. This way, by observing the generated rule set for each table, we could understand to which class attempt (first, second, third or fourth) the rule referred to.

3.3. Retention Heuristics

Our case study concerns a new programme in our university where no student has graduated yet in Information Systems. However, as explained, this is even worse from the University's point of view, because teachers are allocated to teach only few students in a class, those who are following the recommended classes. This brings huge financial losses to the university. Intuitively, we may expect a student to be retained if he starts failing classes and there is no hope to compensate on later semesters or over summer intensive classes. However, since classes offerings vary every year, there is no way to predict the retention of the student using such heuristic. A more viable heuristic was proposed using as basis the recommended programme for a student for every semester (e.g. a student is expected to take the classes MATA68 in his second semester). If in a given semester, a student does not enroll in a recommended class, this might be due to him have failed the pre-requisite classes, or that he just did not want to take that class in a given semester. Thus we verify if there are any pre-requisites to that class that the student have failed. If he failed in at least one, this student is considered to be retained. More specifically, the adopted retention heuristic is as follows:

Retention Heuristic: For every recommended class for a student in a given semester, if the student failed in at least one pre-requisite of a recommended class he didn't enroll, so the student is retained.

After analyzing the preprocessing done, the Apriori algorithm was carried out in order to generate our association rules for each course by each semester, considering also the retained students.

The programme tackled in our case study, i.e. Information Systems, is in its sixth relative semester, however our dataset was set from 2011.1 up to 2012.2 (second relative semester of 2012). Other semesters which do not have a minimal data set required was discarded because no significant results could be obtained.

Table 1. Semester Relative Rules

Rule ID	Rule	Semester Relative
1	mata42dis2=FAIL 22 → retained=t 22 conf:(1) - lift:(1.71)	2
2	matc90dis1=FAIL 37 → retained=t 32 conf:(0.86) - lift:(1.48)	2
3	matd04dis1=APPROVED 16 → matc90dis1=APPROVED 13 conf:(0.81) - lift:(3.91)	2
4	matc94dis1=FAIL 8 → retained=t 8 conf:(1) - lift:(1.28)	3
5	mata02dis2=FAIL 15 → retained=t 14 conf:(0.93) - lift:(1.2)	3

We can observe in Table 1 an example of 5 rules generated by the Apriori algorithm used in our approach. The first part concerns the name of each course and how many students have failed or approved by rule. The *FAIL/APPROVED* parameter depicts if a student was failed or approved in this course. The following data informs if a student is *retained or not* (**t**rue or **f**alse). Afterwards the rule presents each measure analyzed to get significant rules.

4. Results

We used the Apriori association mining algorithm [Agrawal et al. 1994] to mine our rules. Authors in [Agrawal et al. 1993] define an association rule as follows: An association rule is an implication of the form X $\rightarrow$ Y, where X $\subset$ I, Y $\subset$ I and X $\cap$ Y =$\emptyset$. Where I = $\langle i_1, i_2, ...i_n \rangle$ be a set of literals, called items, and D is a set of transactions, where each transaction T is a set of items such that T $\subseteq$ I. A transaction T contains X (a set of some items in I) if X $\subseteq$ T. The rule X $\rightarrow$ Y holds in the transaction set D with confidence c if $c\%$ of transactions in D that contain X also contain Y . The rule X $\rightarrow$ Y has support s in the transaction set D if $s\%$ of transactions in D contain X $\cup$ Y [Agrawal et al. 1993]. In this study, the confidence value was set to 75% and the minimum support was set to 10%, because of the characteristics of our Information Systems programme: i) it is relatively new, it means that no student has already concluded; ii) the number of inflows per year is 45, thus over 3 years, only 135 students have been analyzed.

The lift measure allows evaluating the dependency between a rule precedent and its consequent. In general, this measure expresses how much the resulting becomes frequent when the precedent occurs. For association rules X $\rightarrow$ Y, the lift indicates how much more common it becomes Y when X occurs. The lift is defined as follows: Lift(X $\rightarrow$ Y) = Conf(X $\rightarrow$ Y)/ Sup(Y) when the lift is greater than 1 means that the rule tends to be significant. Therefore, the lift was set to 1.1 in this work [Brin et al. 1997].

Given our retention heuristic, we verified if the student failed a given class previously to analyze his second semester and onwards. Concretely, we analyzed the second and third semesters of each student. The subsequent semesters were not used because there were few students enrolled and taking courses, thus the rules mined were not significant.

Given this, we applied the apriori algorithm in each semester (suport = 10 %, confidence = 75 %, lift = 1.1), generating 24 rules for the second semester and 69 for the third. While the support and confidence may seem relatively low, we must be aware that this programme is relatively young and furthermore has a low dataset size. In order to confirm that this would not bias our results, we used another tool[Webb 2000] to validate our significant rules and measures used. We have obtained the same results thus revalidating our significant rules obtained.

Out of the filtered rules, most were related to when the students were approved in one class would lead to approval in another. Given our hypothesis, this would be indicative that in general given a high confidence and support values, the current organization of the recommended programme workflow is having a positive result. There is also the chance that group of students may be adapting equally the course recommendation and being successful, although unlikely.

We turn our attention now to the rules which had a negative impact in students results, leading to a considerable amount of students failing a class due to another. Again, the second and third semesters are relative, that is, we may have students from different enrollment years being compared, so long that they already concluded the respective relative semester.

4.1. Second relative semester

The first rule (rule 1 in Table 2 depicts that from a set of 22 failed students in Discrete Math (MATA42) for the second time, all of them is retained. This rule validate our heuristic because students that do for the second time the course MATA42 could not enroll in the course Logic Math Introduction (MATC73) neither in Data Structures (MATD04) because of their prerequisite on MATA42. Thus all students that is in the second relative semester following MATA42 for the second time is retained.

Table 2. Semester Relative Rules

Rule ID	Rule	Semester Relative
1	mata42dis2=FAIL 22 → retained=t 22 conf:(1) - lift:(1.71)	2
2	matc90dis1=FAIL 37 → retained=t 32 conf:(0.86) - lift:(1.48)	2
3	matd04dis1=APPROVED 16 → matc90dis1=APPROVED 13 conf:(0.81) - lift:(3.91)	2

The second rule (rule 2 in Table 2) states that from 37 students that followed Digital Circuits and Computer Architecture (MATC90) in the second semester, 32 students is retained. Thus, this rule depicts that 87% of the students that fail in MATC90 course are retained.

Analyzes with domain experts enumerated some causes for this retention values: i) the student is dedicating his time to be approved in courses from the first semester, that he failed; ii) the student cannot conciliate all failed courses with MATC90; iii) as he is already retained, he is not motivated anymore to be approved. Within this study, our main focus is to evaluate the distribution of a programme flow chart. MATC90 does not have any prerequisites, but it has an important number of students that fails. Thus, we can firstly conclude and suggest that a prerequisite be included.

The third rule (rule 3 in Table 2) stated that from 16 students approved on Data Structures (MATD04), 13 students are also approved on Digital Circuits and Computer Architectures (MATC90). 81% of the students that have approved on MATD04, have also approved on MATC90. The majority of the students that do both of these courses for the first time is approved. Thus, as our second conclusion we can state that the organization of both of courses on the programme flow chart of Information Systems is prudent and coherent.

4.2. Third relative semester

Considering the third relative semester, from 8 students that fails Formal Language and Theory (MATC94), all of them is retained, because of MATA42 that is the pre-requisite (rule 4 in Table 3). MATC94 has only a single pre requisite, MATA42. Thus, we can conclude that a single prerequisite is not offering the minimum requirements to be approved. We can suggest, as our third conclusion, that an analysis on all courses that retained the students should be performed so as to incorporate some of those pre requisites into MATC94.

Table 3. Semester Relative Rules

Rule ID	Rule	Semester Relative
4	matc94dis1=FAIL 8 → retained=t 8 conf:(1) - lift:(1.28)	3
5	mata02dis2=FAIL 15 → retained=t 14 conf:(0.93) - lift:(1.2)	3

The fifth rule (rule 5 in Table 3) stated that from all 15 students that were following Calculus A (MATA02) for the second time and have failed, 14 of them were retained. Calculus A is a recommended course for the first semester and it is not a pre-requisite for any course at the second or the third semester. In this way, the fails on MATA02 does not imply in student retention, as stated by our heuristic.

5. Threats of Validity

There are a couple of threats to this work that are important to be highlighted: First, and more important, we are analyzing retention in a programme that is still under implementation. As previously stated, a programme under implementation can be impacted heavily by retained students, since classes may still being planned for the next semesters. Thus, the need for investigating methods of retention in programmes that are still under implementation. We minimize this threat by using our retention heuristic, and the fact that given each class is only offered once a year, a student who failed a class which is pre-requisite to another will be retained. Second, the size of our dataset is relatively small in comparison to other programmes in our university. Third, while variables dichotomization were theoretically validated by a domain expert, the chosen intervals might not work for other programmes. Forth, the lack of more specific class information limits the strength of conclusions. For instance, it is known and observed that students make choices based on which professor is offering the class. Such choices are not taken into account in our work due to privacy reasons.

6. Conclusion and Future Work

Although the adopted retention heuristic may seem too simple to approach the first heuristic, we expect it to approximate it well because of a set of current characteristics of the Information Systems programmes: (1) Every class is only offered once a year, as enrollment only occurs per year, and (2) There are no optative or elective classes being recommended to a student up to the oldest students enrolled in this programme.

In this work, we have tested the hypothesis that the combination of classes that students take on each semester may impact in the chance of them to be retained. In particular, in our case study we created an hypothesis to approximate retention. As we already highlighted over the introduction, the non regular flow of academic student's life

cause some prejudices to the university, especially in programmes under implementation because they must offer the required classes per semester. Having observed a small subset of students that is following the regular recommended flow chart of the Information Systems programme, we have investigated the possibilities of using heuristics to analyze retention in a programmes which is still under implementation.

Although the authors of this work understand that no causal relation can be drawn from observable data, by testing multiple hypothesis in respect to the available student information we can gain further insight on providing solutions to diminish student retention. As our future work, we are interested in extending our analysis by testing two additional hypothesis: Including student related variables and how they influence retention, and also how students social network may affect retention (if at all). We are also planning on extending our analysis to other programmes in our university, and observing which retention patterns hold true within and among the university programmes.

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499.

Barros, R. P. and Mendonça, R. (1998). Conseqüências da repetência sobre o desempenho educacional. *Projeto de Educação Básica para o Nordeste*, Série Estudos(7).

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*, volume 26, pages 255–264. ACM.

Campello, A. d. V. C. and Lins, L. N. (2008). Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior. *XXVIII Encontro Nacional de Engenharia De Produção, RJ, 13p*.

Filho, N. d., Mesquita, F., Marinho, M., Lopes, A. A., Lins, E., and Ribeiro, N. e. a. (2010). Memorial da universidade nova. Technical report, Universidade Federal da Bahia.

Manhães, L. M. B., da Cruz, S. M. S., Macário Costa, R. J., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1.

Manhães, L. M., Cruz, S. M., Costa, R. J., Zavaleta, J., and Zimbrão, G. (2012). Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *VIII Simpósio Brasileiro de Sistemas de Informação (SBSI 2012)*.

Nandeshwar, A., Menzies, T., and Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, 38(12):14984–14996.

REUNI (2012). Reestruturaçao e expansão das universidades brasileiras.

Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução a Mineração de Dados*. Ciência Moderna.

Webb, G. I. (2000). Efficient search for association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–107. ACM.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., and Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8(2010):307–325.

Zhang, Y., Oussena, S., Clark, T., and Kim, H. (2010). Use data mining to improve student retention in higher education - a case study. In *ICEIS (1)*, pages 190–197.