

An Analysis of Degree Curricula through Mining Student Records

Vinicius Gottin¹, Haydée Jiménez¹, Anna Carolina Finamore²,
Marco A. Casanova¹, Antonio L. Furtado¹ and Bernardo P. Nunes^{1,3}

¹Dept. of Informatics, PUC-Rio, Rio de Janeiro, RJ, Brazil

²CEMAT and Dept. of Mathematics, Instituto Superior Técnico, Portugal

³Dept. of Applied Informatics - UNIRIO, Rio de Janeiro, RJ, Brazil
{vgottin, hjimenez, casanova, furtado, bnunes}@inf.puc-rio.br, anna.couto@ist.utl.pt

Abstract—Higher Education Institutions store a sizable amount of data, including student records and the structure of a degree curriculum. This paper focuses on the problem of identifying how closely students follow the recommended order of the courses in a degree curriculum, and to what extent their performance is affected by the order they actually adopt. It addresses this problem by applying techniques to mine frequent itemsets to student records. The paper illustrates the application of the techniques for a case study involving over 60,000 student records in two undergraduate degrees at a Brazilian University.

Keywords—degree curriculum structure; academic analytics; frequent itemsets

I. INTRODUCTION

One of the major challenges faced by Higher Education Institutions (HEIs) is to ensure that students succeed, which could be broadly defined as students' retention and graduation. An analysis of student records and degree curricula can help HEIs minimize student dropout rates, and meet the demands of the students.

In this paper, we investigate the adequacy of degree curricula by applying techniques to mine frequent itemsets [10, 14] to student records, in a scenario where students have some freedom to choose the courses they enroll in. We address specific questions related to degree curricula: in a given semester, which sets of courses students typically enroll in (or typically fail)? Are these sets consistent with the set of courses recommended for the given semester? How much the courses recommended for a given semester affect student performance?

The major contribution of this paper is two-fold. First, we argue that the concept of *frequent itemset* can be adopted to capture the intuitive notions of “typically enroll in” (and “typically fail”). An *item* in this case is a course a student enrolls in (or fails), a *transaction* or a “*basket*” is the set of courses a student enrolls in (or fails) in a given semester. Hence, a frequent itemset will be a set of courses students frequently enroll in (or fail) in a given semester. We then show how the techniques to mine frequent itemsets can be applied to student records to answer the questions we address, whereas other standard process mining techniques do not seem to be useful in this context. Furthermore, we note that the approach we propose goes beyond the mere computation of statistics about student enrollment (or failure) in *individual* courses, since we mine frequent *sets* of courses.

Secondly, we apply the proposed frequent itemset techniques to a representative set of student records of a Brazilian university to investigate the adequacy of the degree curricula. To illustrate the analysis, we discuss in detail two such degrees, Law and Architecture, and show that they exhibited different profiles.

The remainder of the paper proceeds as follows. Section II briefly reviews the literature in analytics in education and related fields. Section III describes the use case scenario adopted throughout this paper. Section IV presents an in-depth analysis through basic statistics and a frequent itemset mining algorithm to investigate the adequacy of the degree curricula and student performance at the selected university. Finally, Section V highlights the contributions of the paper and proposes future works.

II. RELATED WORK

Analytics in Education is an emerging topic that helps academic administrators devise strategies to attract new students, understand student retention patterns, and redesign degree curricula, among other aspects. According to Norris et al. [12], the term *analytics* encompasses data analysis and assessment that lead to improvements and performance tracking in HEIs.

There are many studies about Analytics in Education on topics such as Learning Analytics, Academic Analytics and Predictive Analytics. For instance, in the Academic Analytics field, Goldstein et al. [13] conducted an extensive survey to identify how HEIs exploit institutional data to support decision making. As for the Learning Analytics field, researchers have been interested in investigating student dropout [15, 16] and retention [1, 17, 18, 19], students' profile identification [8, 20], and student performance prediction [9, 7, 21]. Unlike our approach, these studies mainly focus in machine learning techniques such as clustering, classification, and regression rather than mining techniques.

Mining techniques may also be useful for supporting Academic Analytics. For instance, process mining techniques [2] extract knowledge based on event data stored in information systems to automatically create and discover process models within an organization. Although such techniques may help HEIs in several ways, in the problem under investigation in this paper, process mining algorithms - such as alpha, heuristic, fuzzy and inductive, developed in software tools, such as ProM [11] - proved not to be useful since the order in which students enroll in courses has to

respect their pre-requisites, which are known a priori, and therefore such tools uncover no new knowledge. To understand the effect of the order of the courses dictated by the degree curriculum on students' performance, we resorted to data visualization techniques [3,4] and algorithms to mine frequent itemsets [10,14].

Indeed, our approach uses frequent itemset mining algorithms [10], where each itemset is a set of courses in which a student enrolled, in a given semester. This approach proved to be convenient to characterize students' enrollments. It took into consideration the student's completion status in each academic term. To our knowledge, there have been no such studies based on frequent itemset mining. Although, for a different purpose, Huang et al. [22] used frequent itemsets to find learner behavior patterns in an online degree.

III. CASE STUDY

The datasets used for the case study were provided by the academic system of the Brazilian university on hand in multiple formats, spanning from 1996 to 2016. In this paper, we selected the 2002 Architecture degree curriculum and the 2008 Law degree curriculum – Litigation Emphasis (see Table 1), since they are representative curricula of two fairly different undergraduate degrees. They are identified by a mnemonic code: ARC – Architecture; LL – Law, Litigation emphasis. From this point on, we will simply refer to degrees rather than degree curricula.

The final dataset consists of 60,658 records of 1,661 distinct students enrolled in the two selected degrees. Each student record consists of the enrollment (and status) of a student in an offering of a course of a degree. Among the 31 features in the dataset, we highlight the most important ones for the analysis and visualizations produced in this work. They are: student ID; degree code; course code; year of enrollment; semester of enrollment; group to which a course belongs (if any); the recommended semester of a course; the effective semester in which the student enrolled in a course; the student's status in a course (*canceled*, *passed* or *failed*); and student's academic status (*enrolled*, *abandoned* or *graduated*).

Some data processing tasks were performed to prepare the data for further analysis. For instance, the effective semester of each student's enrollment in each course was calculated as the difference, in semesters, between the current academic term at the time of enrollment, and the student's first enrollment in the degree, e.g., if a student first enrolls in the degree in the 1st semester of 2015 and enrolls in a course in the 1st semester of 2016, the effective semester of the enrollment in that course is the 3rd, that is, the student takes that course in his 3rd academic term.

The effective semester is useful to calculate the advance or delay of students' enrollment in courses, defined as the difference between the effective semester and the recommended semester of the course, as specified in the degree recommended order. Following our previous example, if the student enrolls, in his 3rd academic semester, in a course that is recommended for the 1st semester, he is delaying enrollment in that course by 2 semesters. Conversely, if the recommended semester is the 5th semester, the student is

Table 1. Data Summary

Semester	ARC (2002)		LL (2008)	
	#Students	#Records	#Students	#Records
1	1,199	6,885	344	2,804
2	1,086	5,575	344	2,807
3	1,002	5,606	341	2,408
4	922	4,321	344	2,730
5	871	4,449	339	2,543
6	810	4,076	334	2,256
7	720	3,654	264	1,065
8	652	3,279	249	644
9	590	2,621	282	551
10	536	1,929	272	455
TOTALS		42,395		18,263

taking that course 2 semesters in advance of the recommended semester.

Finally, we stress that the pre-requisites force students to take courses in the desired order. However, they do not guarantee that students will take the courses in the appropriate semester, which is an issue worthy of the academic administrators' attention.

IV. EMPIRICAL ANALYSIS

A. Basic Statistical Analysis

The performance of a student is typically measured by the grades he obtains in the courses he enrolls in and by his average grade. In the university in question, a student fails to pass a course when his average grade is less than 5.0 (the top grade is 10.0). As anticipated in the previous section, we classify students into 3 broad groups: students that graduated; students that abandoned the degree for some reason; and students currently enrolled.

A very basic question is therefore:

Q₁: How the 3 student groups differ in terms of students' performance?

To illustrate how to answer this question, Figure 1 shows a normalized histogram of the average grades for the ARC degree. Note that, for the 3 groups, grades below 5.0 are less frequent, with the exception of grade 0.0, which is typically given to students that do not cancel their enrollment, but forfeit the course, either by grade or by number of absences. Also, note that the ratio between graduated and abandoned students is 1.19.

With a general idea of the profile of the students in the degree, we proceeded to compare the abandoned and graduated students with the current enrolled students. For the ARC degree, Figure 1 indicates the pattern displayed for the currently enrolled students is quite similar to that of the graduated students, which is a good indication in terms of retention. This is not true, however, for the LL degree, since students' performance in this case is very similar for all the 3 groups, and the ratio between graduated and abandoned is 1.88.

Another interesting question is:

Q₂: In which semester students more often abandon the degree?

To answer this question for the ARC degree, Table 2 shows the percentage of number of students that abandoned the degree in each semester with respect to the total number

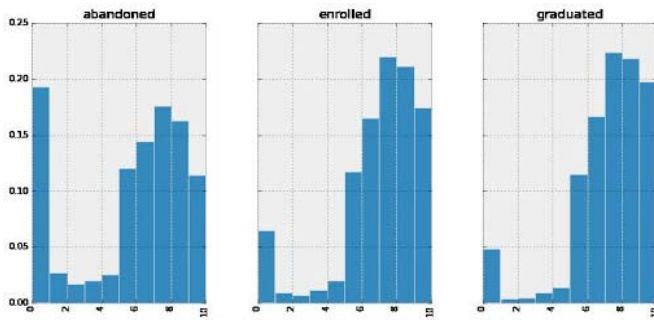


Figure 1. Normalized histogram of ARC students' grades.

Table 2: Dropout rate for each semester of the ARC degree.

Semester	1	2	3	4	5	6	7	8	9	10
%Abandoned	34.6	13.7	12.4	7.5	6.9	8.5	5.9	5.2	1.6	3.6

of students that abandoned the degree in the period under investigation. Note that the highest abandonment percentage occurs in the first semester. There are at least two possible reasons for this: (i) courses taken in the first semester do not meet students' expectations and consequently do not motivate them; or (ii) the admission exams do not guarantee that the students are prepared to take those courses. In both situations, a reorganization of the curriculum or of the courses' syllabus should be considered, in order to decrease abandonment rates in the first semester.

A third question is:

Q_3 : For each course, how much students advance or delay enrolling in the course, with respect to the recommended semester, vis-à-vis the approval rates?

To answer this more complex question, we adopted the visualization strategy illustrated in Figures 2 and 3. The visualization consists of the proportion of enrollments in each course recommended for a given semester. The size of the square in each cell gives the percentage of students that advanced enrollment in that course (columns marked with a negative integer) or delayed enrollment (columns marked with a positive integer). The color also evidences delay or advance, according to the legend.

Figure 2 shows an example of this visualization for the ARC degree. In this case, students neither advance nor delay courses as often, except for 3 courses (2 on the left and 1 on the right of the middle column). For the LL degree, Figure 3 indicates that there are 3 courses for which students very frequently advance enrollment in the 7th, 8th and 9th semesters – a fact that should be further investigated, as it suggests that changing the order of the courses of the current LL degree curriculum might better suit the students' demands.

B. Frequent Itemsets Analysis

In this section, we focus on two further questions:

Q_4 : For a given semester, how closely the courses students typically enroll in match the set of courses the degree curriculum recommends for the semester?

Q_5 : For a given semester, in which courses, individually or in group, students typically fail?



Figure 2. Analysis of student enrollment and performance in the ARC degree (first 3 semesters).



Figure 3. Analysis of student enrollment and performance in the LL degree (final semesters).

Note that in the previous section, we considered statistics about the courses independently of each other, which are not sufficient to address these questions, since they involve sets of courses.

A first answer to these questions would be to create an index that, for a given semester, directly compares the set of courses students enroll in (or fail to pass) with the set of courses recommended for the semester. More precisely, let ST be a given set of students enrolled in a degree D over a period of time T , measured in semesters. The *degree-semester adherence index*, denoted $I_{D,S}$, measures how much the students in ST followed the recommended set of courses for degree D at a given semester S in T . It is defined as:

$$I_{D,S} = \frac{1}{n} \sum_{i=1}^n \frac{|E_i \cap R|}{|E_i \cup R|}$$

where: n is the number of students in ST enrolled in D in semester S ; E_i is the set of courses student i in ST enrolled in semester S ; R is the set of courses recommended for semester S of D ; and the fraction in the summation is the Jaccard similarity between E_i and R [10]. Note that $I_{D,S} \in [0,1]$, where $I_{D,S} = 0$ iff no student enrolls in any of the recommended courses for semester S of D , and $I_{D,S} = 1$ iff all students enroll in exactly the recommended courses. The *degree adherence index* would then be defined as the average of degree-semester adherence indices for the semesters of the degree.

Figure 4 illustrates the degree-semester adherence index. For both degrees, a sharp drop is observed in the 7th semester, which means that in this semester students are not following the recommended courses. Figure 3 corroborates this behavior for the LL degree, and the conclusion is that the low index value for this semester is due to the fact that students very often advance enrollment in the (single) course, JUR1836,

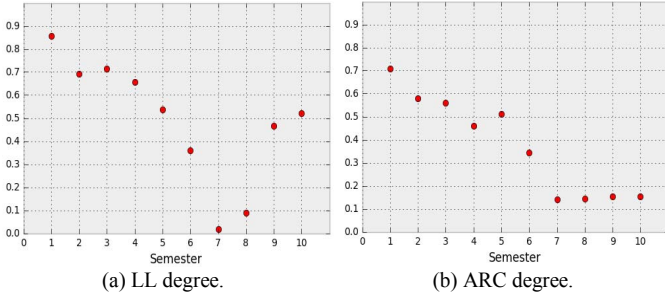


Figure 4. Degree-semester adherence index.

suggested for the 7th semester. Another difference between the 2 degrees is the behavior in the last 2 semesters: while students frequently follow the recommended courses for the LL degree, indicated by a high index for these semesters, this is not true for the ARC degree. One possible explanation would be that ARC students take more than the 10 recommended semesters to graduate.

The degree adherence indices are 0.49 for LL and 0.37 for ARC, both of which are fairly low, and therefore suggest that the academic administrators should indeed rethink the current recommended curricula.

We now provide more detailed answers to questions Q_4 and Q_5 by applying techniques to mine frequent itemsets to the student records. We recall that mining frequent itemsets is based on finding the relative frequency with which two or more objects of interest co-occur in a dataset [5].

First, for each of the 3 course statuses in a semester – *enrolled*, *approved* or *failed* – we applied an implementation of the Apriori algorithm [6]. However, since the algorithm may return too many sets of courses, together with their frequency, a direct analysis of the output would not help draw conclusions about the set of courses most often followed by the students.

Thus, to tackle this issue, we run the hierarchical agglomerative clustering, considering the Ward’s linkage method, for each status and each semester. Based on the resulting dendrogram, and aided by the academic specialists, we defined the number of clusters. Next, for each cluster, we kept only the maximal sets of courses. For example, in Figure 7, both {ARQ1101, MAT1071} and {ARQ1000} belong to the same cluster (identified by horizontal dashed lines) but none contains the other. So, both sets are kept.

To illustrate how question Q_4 can be answered, consider the students enrolled in the first semester of the ARC degree. Figure 5 shows the maximal sets for each cluster, that is, the most frequent sets of courses recommended for the first semester taken by students of the ARC degree. Each bar corresponds to a set and its size portrays the number of times students enrolled in all courses in the set. The upper bar represents the most frequent set of courses and the bottom bar (hatched) represents the set of courses recommended by the curriculum. Observe that these two sets differ in only one course, CRE1100. The second most frequent set includes a new course, ART1028, which is not recommended for the first semester, and does not include CRE1100. This leads us to conclude that CRE1100 could be delayed in the curriculum.

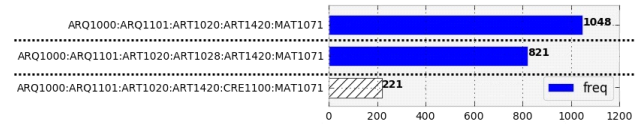


Figure 5. Frequent itemsets, enrolled stud., 1st semester, ARC degree – highlighting the recommended set of courses (hatched).

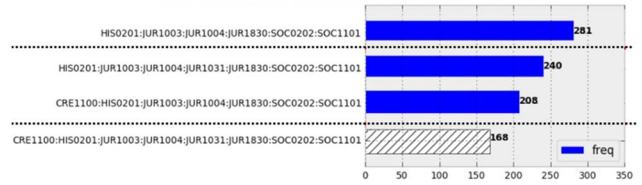


Figure 6. Frequent itemsets, enrolled students, 1st semester, LL degree – highlighting the recommended set of courses (hatched).

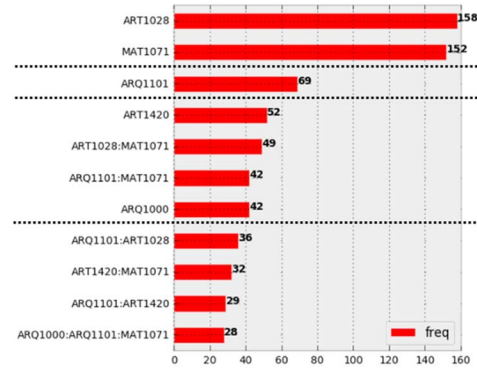


Figure 7. Frequent itemsets, failed students, 1st semester, ARC degree.

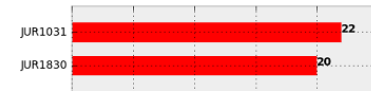


Figure 8. Frequent itemsets, failed students, 1st semester, LL degree.

Figure 6 shows the same data for the LL degree. Again, the bottom bar corresponds to the set of courses recommended for the first semester. Note that the number of sets of courses in which students frequently enroll in the first semester of the LL degree largely exceeds that of the ARC degree.

Question Q_5 can be answered using a similar strategy. Figure 7 shows the courses students frequently failed in the 1st semester of the ARC degree. The singletons indicate the courses that students frequently failed. The top failed course, ART1028, does not belong to the set of courses recommended for the 1st semester, which can be checked by observing Figure 2. By a further analysis, we found that this course is really not problematic, since its approval rate is in fact much higher than the failure rate (the ratio between students that are approved to students that failed is about 4.43). Note that the sets {ART1028, MAT1071} and {ARQ1101, MAT1071} are also frequent, that is, students also frequently fail in both courses in these sets when they enroll in them in the first semester. In addition, note that this fact would not be uncovered by statistics of student failure for individual courses.

Figure 8 shows that the frequency of failed students for the first semester of the LL degree is fairly low, being critical in just two individual courses, JUR1031 and JUR1830.

ACKNOWLEDGMENT

This work was funded by CNPq, under grants 303332/2013-1, 442338/2014-7 and 557128/2009-9, and by FAPERJ, under grants E-26-170028-2008 and E-26/201.337/2014.

REFERENCES

- [1] E.J. Lauria, J.D. Baron, M. Devireddy, V. Sundararaju, and S.M. Jayaprakash. Mining academic data to improve college student retention: An open source perspective. Proc. 2nd Int'l. Conf. on Learning Analytics and Knowledge, Vancouver, 2012, pp. 139-142.
- [2] W.M. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, London, 2011.
- [3] J. J. van Wijk. The Value of Visualization. *IEEE Visualization*, 2005.
- [4] J. Heer, M. Bostok, and V. Ogievetsky. A Tour through the Visualization Zoo. *Comm. of the ACM*, vol. 53, no. 6, pp. 56-67, 2010.
- [5] M.J. Zaki and W. Meira Jr.. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.
- [6] R.J. Bayardo Jr., "Efficiently Mining Long Patterns from Databases," Proc. SIGMOD Int'l. Conf. on Management of Data, pp. 85-93, 1998.
- [7] B.K. Baradwaj and S. Pal. Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 6, pp. 63-69, 2011.
- [8] S. Parack, Z. Zahid and F. Merchant. Application of data mining in educational databases for predicting academic trends and patterns. Proc. IEEE Int'l. Conf. on Tech. Enhanced Education, pp. 1-4, 2012.
- [9] M.M. Abu Tair and A.M. El-Halees. Mining educational data to improve students' performance: a case study. *Int'l. J. of Information and Comm. Technology Research*, Vol. 2, No. 2, pp. 140-146, 2012.
- [10] J. Leskovec, A. Rajaraman, J.D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.
- [11] B.F. van Dongen, A.K.A. de Medeiros, H.M.W. Verbeek, A.J.M.M. Weijters, and W.M. van der Aalst, The ProM framework: A new era in process mining tool support. Proc. Int'l. Conf. on Application and Theory of Petri Nets, pp. 444-454, 2005.
- [12] D. Norris, L. Baer, M. Offerman. A national agenda for action analytics. Proc. Nat. Symposium on Action Analytics, pp. 21-23, 2009.
- [13] P.J. Goldstein, R.N. Katz. Academic analytics: The uses of management information and technology in higher education. 2005
- [14] F.A. Amorim, B.P. Nunes, G.R. Lopes, M.A. Casanova, MFI-TransSW+: Efficiently Mining Frequent Itemsets in Clickstreams. Proc. 17th Int'l. Conf. on Elect. Comm. and Web Technologies, 2016.
- [15] L. Agnihotri, A. Ott. Building a Student At-Risk Model: An End-to-End Perspective From User to Data Scientist. *Educ. Data Mining* 2014.
- [16] E. Yukselturk, S. Ozekes, Y.K. Türel. Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning*, 2014.
- [17] M. Bogard, T. Helbig, G. Huff, C. James. A comparison of empirical odels for predicting student retention. White paper. Office of Institutional Research, Western Kentucky University, 2011.
- [18] C.E. Calvert. Developing a model and applications for probabilities of student success: a case study of predictive analytics. *Open Learning: The Journal of Open, Distance and e-Learning*, 2014.
- [19] E. Aguiar, N.V. Chawla, J. Brockman, G.A. Ambrose, V. Goodrich. Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. Proc. 4th Int'l Conf. on Learning Analytics and Knowledge, pp. 103-112, 2014.
- [20] R. Phillips, D. Maor, G. Preston, W. Cumming-Potvin. Exploring learning analytics as indicators of study behaviour. Proc. EdMedia: World Conf. on Educ. Media and Tech., pp. 2861-2867, 2012.
- [21] A.B. Ahmed, I.S. Elaraby. Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*. 2014.
- [22] J.P.Huang, S.J. Chen, H.C. Kuo. An efficient incremental mining algorithm-QSD. *Intelligent Data Analysis*. 2007.

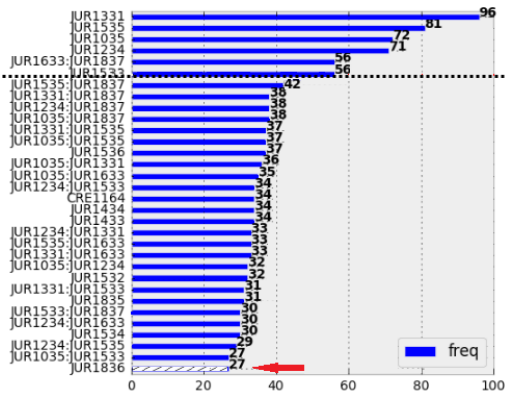


Figure 9. Frequent itemsets, enrolled students, 7th semester, LL degree – highlighting the recommended set of courses (hatched).

Figure 9 shows the frequent itemsets for the 7th semester for the LL degree, which has only 1 recommended course, JUR1836 (indicated by the red arrow). Its low frequency signals that students rarely enroll in JUR1836 at the recommended semester. As already mentioned, Figure 3 in fact shows that students more frequently enroll in JUR1836 in advance, in the 5th or the 6th semesters.

Furthermore, Figure 4 shows that the 7th semester of the LL degree has a very low semester adherence index, which confirms that students frequently do not care to follow the set of recommended courses for the 7th semester, which in this case just consists of JUR1836.

V. CONCLUSIONS

In this paper, we applied frequent itemset mining techniques and visualization strategies to analyze degree curricula. We summarized the current efforts towards analyzing the academic domain at the university under investigation. The case study dataset comprised over 60,000 records of nearly 1,700 students.

The results show that the analysis of academic domains as processes is a promising approach. However, we stress that process mining software tools, such as ProM [11], were not helpful, for the reasons explained in Section 2. We used instead techniques for mining frequent itemsets. In our case study, these techniques were fundamental to reveal when students do not follow the recommended order of courses, as well as to identify which courses were most frequently delayed or advanced by students.

The conclusions drawn from the application of these frequent itemset techniques also guided the production of custom visualization strategies. These custom visualizations purport to allow academic administrators to identify, for each course, the impact of delaying or advancing a course on the student's approval and performance ratios. Having access to data analysis reports of this kind, they are better prepared to reason about the quality of the recommended curriculum.

Our long-term goal, following a prescriptive orientation, is to incorporate such methods in a tool to help academic administrators compose recommended curricula in a way that effectively improves student performance.