Communications of the Association for Information Systems

Volume 16 Paper

September 2005

Applying Data Mining to Scheduling Courses at a University

Wayne Smith
Clarement Gradute University, wayne.smith@cgu.edu

Recommended Citation
Smith, W. (2005). Applying Data Mining to Scheduling Courses at a University. Communications of the Association for Information Systems, 16, pp-pp. https://doi.org/10.17705/1CAIS.01623

Communications of the Association for Information Systems

APPLYING DATA MINING TO SCHEDULING COURSES AT A UNIVERSITY

Wayne Smith
School of Information Systems and Technology
Claremont Graduate University
wayne.smith@cgu.edu

ABSTRACT

Scheduling courses ("timetabling") at a University is a persistent challenge. Allocating course-sections to prescribed "time slots" for courses requires advanced quantitative techniques, such as goal programming, and collecting a large amount of multi-criteria data at least six to eight months in advance of a semester. This study takes an alternate approach. It demonstrates the feasibility of applying the principles of data mining. Specifically it uses association rules to evaluate a non-standard ("aberrant") timetabling pilot study undertaken in one College at a University. The results indicate that 1), inductive methods are indeed applicable, 2), both summary and detailed results can be understood by key decision-makers, and 3), straightforward, repeatable SQL queries can be used as the chief analytical technique on a recurring basis. In addition, this study was one of the first empirical studies to provide an accurate measure of the discernable, but negligible, scheduling exclusionary effects that may impact course availability and diversity negatively.

Keywords: timetabling, data mining, course scheduling, decision-support systems, institutional research, University administration

I. INTRODUCTION

> *"Although progress has been made, solving large instances [of course scheduling problems] is difficult."* [Mooney, et. al. ,1996, p. 369]

Scheduling courses at a University is a recurring and complex operational activity [Burke, et. al., 1997]. Course scheduling is often done in an environment

- where an initial schedule needs to be built many months in advance of the term,
- subject to a large number of operational constraints,
- without major computational assistance, and
- instantiated by individuals with incomplete or ephemeral information.

Course scheduling is referred to as a "timetabling" problem in the education literature (e.g.,, Kumar, [2004] or Tripathy, [1984]) and as an "assignment" problem in the operations research literature (e.g., Boronico [2000], Hertz and Robert [1998], or Stallert, [1997]). Generally, solutions to these types of large-scale, data- and computationally-intensive problems in higher education

are formulated and implemented with a variety of quantitative methods, typically goal programming [Cheng, 1993]. This approach requires both a strong organizational commitment to a structured and mathematically-based paradigm and an infrastructure to capture the data needed to populate the constraints and variables in any given model. The first requirement may be feasible but impractical at many universities, and the second requirement is not feasible at all in a timely and accurate manner across many federated departments.

By definition, the use of a formal goal programming method would require many known parameters. Even, or especially, for universities that do not use a formal goal programming method to schedule classes, a rich understanding of the quantity and dispersion of critical parameters is helpful. An initial step, but one that is difficult to do in large problem spaces with correspondingly large databases, is to analyze the data to illuminate important bivariate relationships in course-section enrollments, especially the subtle conflicts between required courses and allocated time-slots. Subsequently, analysts can evaluate the resulting course-section timetabling conflicts and changes in student enrollment patterns in detail, often by traditional managerial processes.

This paper describes the application of data warehouse and data mining principles to generate appropriate measures and describe relevant patterns of one such pilot study. Inductive methods that elicit patterns from large datasets are core characteristics of data mining approaches [Mannila, 2000]. However, feasibility is a necessary, but insufficient condition for adoption and use of innovations [Rogers, 2003]. Many decision-makers may not be familiar with data mining methods, much less trust the theoretical basis for the resulting inferences that are generated inductively. For a difficult problem such as timetabling, a related issue is instantiating a parsimonious solution (i.e., exhaustively complete and elegantly compact) that is both viewed as "successful" and engenders persistent end-user technology use [Delone and McLean, 1992]. To the extent that practitioners feel that they can understand the methods and repeat them with little or no additional training and with the use of existing toolsets, the basis for a new, recurring, and extendable analytic technique is formed. Finally, confronting deeply-held, but potentially incorrect, beliefs about the characteristics embodied in a large organization is important, but difficult. IT-based solutions can assist in that regard, especially in prototyping responses [Benjamin, 1993] from the many stakeholders involved.

II. CONTEXT

The pilot study was undertaken in the College of Business & Economics at California State University, Northridge; an urban University located in Southern California. The University is one of the 25 largest in the country, enrolling approximately 33,000 students in fall 2003. The College enrolled approximately 5,500 students in that semester. The College offers primarily upper-division, professional courses for business undergraduate students. The University enrolls many part-time and commuter students, and somewhat disproportionately so in this particular College.

Course schedules at this University are developed in a relatively decentralized process by chairs at the academic-department level. Except for sections in large-lecture halls, each chair is given an allocation of classrooms, all of which accommodate approximately 40 students. All academic decision-makers agree that a "good" schedule is one that maximizes the likelihood that students are able to enroll in needed or desired courses. While course scheduling is an important issue for all students, it is especially important for part-time students [Keast, 1998].

For Fall, 2003, the University scheduled 4,758 sections of 2,000 courses. The pilot College scheduled 314 sections of 102 courses. Of those 314 sections, the pilot College scheduled 27 "aberrant" sections during "prime-time." "Aberrant" sections are defined as those sections that were either scheduled for 1.5 hours on Monday and Wednesday or for 3.0 hours on Friday (traditionally, they would be scheduled for 1 hour only on Monday, Wednesday, and Friday or for 1.5 hours only on Tuesday and Thursday). "Prime-time" is defined as Monday through Friday

from 8am to 2pm. Courses in the College consist almost entirely of 3-unit courses. Given these parameters, Chairs scheduling aberrant course-sections using alternate time-slots can anticipate generating unavoidable scheduling overlaps with other course-sections both within this College and elsewhere in the University at large.

As might be typical at a large, urban, commuter institution, students are not enrolled in "lock-step" programs in general, and certainly not within this one College. Among the many matriculation issues that are beyond the scope of this paper, it is not uncommon for "seniors" to take "junior-level" courses or even "general education" courses in their senior year. Some of these courses might be in different Departments within the same College, or, worse, in different Colleges within the University. Chairs no doubt receive good information regarding scheduling in general and scheduling Department sections and courses specifically. However, the scheduling "problem space" may simply be too large for even an experienced Chair to have enough good information be able to discern potential scheduling conflicts at a granular enough level to avoid inadvertently increasing the number of scheduling conflicts and potentially limiting a student's choices. The extreme worst case might be an aberrant schedule that absolutely precludes only a single student from enrolling in a course that is required in a program. Note also that even without a single aberrantly scheduled course or section of a course, information about scheduling relationships and patterns is needed to avoid conflicts. The rich scheduling conflict information can be used not only to design alternate schedules parsimoniously, but also audit such schedules empirically on a recurring basis. For example, in addition to the formal, deliberate aberrant scheduling used by this pilot study, aberrant schedules could occur without the explicit knowledge of the University (if the allocation of course-sections into prescribed time-slots only is not enforced via a manual or automated process).

The institution is not looking to replace one set of human schedulers with another or even replace the human schedulers with some type of expert system. The Chairs and many other key decision-makers at the College-level and University-level just need better information about the theoretical or empirical scheduling conflicts. Deriving the theoretical conflicts (such as aggregating the 0.5 hour conflicts between sections that are regularly scheduled from 8am-9am and sections that are aberrantly scheduled from 8am-9:30am) is intuitive to the decision-makers. Moreover, the decision-support, data processing tasks are relatively simple. However, deriving the empirical conflicts (such as summarizing the actual, multivariate patterns of section or course enrollments by students) may be intuitive to most decision-makers, but the decision-support, data processing tasks are decidedly non-trivial. Moreover, while the number of empirical (actual) conflicts can be no higher than the number of theoretical (hypothetical maximum) conflicts, the number of empirical conflicts may be significantly less than the number of theoretical conflicts. In fact, with good information about the theory of enrollment patterns at the course-level and good design practice regarding course offerings at the section-level (where day/time conflicts occur), the number of empirical conflicts can be minimized, even beyond the reduction in empirical conflicts due just to stochastic dynamics.

Putting aside the traditional challenges of organizational adoption of new methodologies for a moment, the primary technological challenges in this situation are:

1. to find the relevant course associations and related interactions, which in turn requires generating the exhaustive pairs of courses taken together, and

2. to generate an intuitive framework and visual format for the Chairs, Deans, Provost's staff and others to understand and use the resulting information actively.

III. RELATED WORK

The published literature on University timetabling typically describes environments where a mathematical procedure, typically a goal programming or heuristic-search process, is executed before the semester begins. Techniques that are used and studied include integer

programming [Tripathy, 1984], "backward scheduling logic" [Cox and Jesse, 1981], *tabu* search algorithms [Hertz, 1991], and many ad hoc, iterative mechanisms [Ferland and Fleurent, 1994; Badri, 1996]. Due to the large number of dimensions and geometric scaling for even small problems, University timetabling attracts the interest of researchers from several disciplines. Among the advanced mathematical techniques used are network theory [Dinkel, et. al., 1989], graph theory [Yu and Yang, 1993; Kiaer and Yellon, 1992], genetic algorithms [Burke, et. al., 1993], and discrete event simulation [Boronico, 2000].

Empirical results from studies of scheduling solely within one College at a U.S. University are informative. Mooney, et. al. [1996] found that an objective of a small number of scheduling conflicts for a few courses is actually preferable to strenuously trying to keep the average number of conflicts small but allowing an arbitrary worst case schedule. Further, Mooney, et. al. [1996, p. 377] found that even after many incremental improvements in the optimization model, "serious challenges…in the areas of…preferences, fairness, and robustness" Still exist. Badri, et. al. [1998, p. 304] found that the "complex utility functions could limit… application when used on a practical, recurring basis…" As an example of the complexity of a typical timetabling problem, Badri, et. al.'s [1998, p. 313] proposed model at the institution he studied "…consists of 252 decision variables, 66 goal constraints and [a total of 167] system constraints."

Even when a well-understood model can be formulated, course scheduling can require factoring a larger problem into at least two smaller sub-problems that are then solved sequentially [Hertz and Robert, 1998]. This multi-stage approach is seen as needed even when scheduling a single College within a University [Stallaert, 1997]. As recently as 2002, one College at a University identified course scheduling as "…a major problem for the school…[and]…the root cause of [other, major logistical problems]" [Hinkin and Thompson, 2002; pg. 1]. Fundamental variables that have to be managed by a department chair manually (even if they are not globally applied), such as faculty preferences for consecutive classes, were not included in prior, but ostensibly comprehensive, models [Hinkin and Thompson, 2002].

Little formal *post hoc* analysis of large-scale aberrant scheduling, much less analytics employing the rigor and relevance of the emerging techniques of data mining, has been done in practice at most universities [Kehoe, 2004]. Published data mining results in the field of University institutional research began to appear in the last two years. Much of the work appears to originate with a relatively few institutional research professionals. Data mining was used to study student persistence [Willett, 2003], student learning outcomes [Juan, 2002], and admission yields [Chang, 2003].

The principles of data mining are identified as the "next revolution in institutional research", and moreover, data mining "…has the potential to shift the institutional research function from a retrospective accounting function to a prospective management function" [Kumar, 2004].

IV. RESEARCH QUESTION

The central research question in this study is:

> Is a data-mining approach to analyzing the consequences of permitting one College to schedule some of its course-sections in time-slots that are inconsistent with established, prescribed time-slots, feasible?

As used here, the term feasible refers to a solution that is technically achievable, methodologically sound, and understandable by decision-makers. A related question is, assuming a feasible solution can be developed and implemented for this pilot study, what is the simplest technique that be used to facilitate future replication by existing institutional staff with existing tools on a recurring basis? Chairs, Deans, and other key academic managers want answers about the consequences of a particular aberrant schedule. The intent of this study was not to answer each and every broad functional question with absolute precision, but rather to

demonstrate a technology prototype that can be used as a foundation to answer functional and emergent questions in an accurate and iterative manner.

Paramount among the functional questions of interest is whether an aberrant schedule during prime-time will lead to an unacceptable increase in the number of actual course-section scheduling overlap intersections (i.e., timetabling exclusions). Any such increase in exclusions potentially decreases all or in part, a student's ability to enroll in required or desired courses. Less visible, but still of keen interest to decision-makers, is the degree to which a material increase in scheduling exclusions, if any, leads to an inadvertent change of student-course enrollment patterns. In addition to the goal of not reducing the quantity of courses available to a student, the University is interested in not reducing the variety of courses available to a student as well, at least not a reduction in variety caused by an aberrant scheduling model. Questions such as these are important because, as with any quality institution of higher education, the University actively strives to

1. minimize impediments to the length of time to graduation,
2. maximize use of existing classroom space,
3. meet expressed student demand for course-section availability, and
4. deploy faculty resources efficiently and productively.

V.METHODS

Although many questions by decision-makers arise in the process of course scheduling, this research focuses on the feasibility of answering a few key, functional (i.e., operational) questions that can be used to augment existing knowledge and traditional univariate reports. For the purpose of this paper, the following functional question is illustrative of the data-mining approach used:

1. Which combinations of courses do students take together and
2. were those combinations different for students enrolled in at least one aberrantly-scheduled course-section?

In the language of data mining, the challenge is to describe the differences, if any, between the dispersion groups (or clusters) that form naturally from the actual student enrollment patterns.

Note that traditional statistical inference techniques are difficult to employ in answering this (not atypical) functional question.

- First, the functional question is less about the distributions of each course and more about the relationship between at least two courses. To answer this question requires, at a minimum, a transformation of transactional data to generate an exhaustive list of bivariate course pairs. Note that even if the pairs are generated, one cannot say without a great deal of subjectivity that one course is the "independent" variable and the other course is the "dependent" variable.

- Second, course data is at the nominal (categorical) level. In this problem context, few interval-ratio variables exist from which to use traditional techniques such as correlation or t-tests.

- Third, it may be difficult for a lay analyst, much less a number of decision makers, to interpret the results of advanced analytical techniques such as homogeneity analysis (correspondence analysis) and predictive analytical procedures found in commercial software tools such as SPSS (SPSS, 2004).

Association rule mining is a non-parametric, data mining technique that generates relevant patterns of association between two or more "itemsets". The term itemsets is used because the derived associations are between sets of observations rather than groups of variables. Note that while SQL-92 defines set concepts such as UNION, INTERSECTION, and EXCEPT (the MINUS operator in Oracle), the actual implementations of these relational algebra concepts varies widely

among database providers. A typical association rule is represented in general as "A → B". One of the more common applications of association rule mining is "market basket" analysis. Such an association rule in Marketing might appear as "beer → diapers". A → B is not to be interpreted as a causal relationship. Buying beer might not lead to buying diapers (or vice-a-verse) as a generalization, but buying beer and diapers together might occur more frequently than other combinations of purchases and therefore inform the Marketing staff as to the placement of such items within a store, for example.

The association rule "beer → diapers" is simply a pattern of association. In any given problem space, it should be clear that there can be many candidate association rules. An association rule is determined to be relevant if it meets two minimum thresholds of "interestingness"—"support' and "confidence" (Han and Kamber, 2001 p. 147). "Support" measures the proportion of individual record patterns (for example, in our problem, pairs of sections or pairs of courses) to the total number of records (e.g., total enrollments). "Confidence" measures the "strength" of the relationship between each individual record (e.g., course "ACCT 220" and course "BLAW 280" as fraction of all "ACCT 220" enrollments). "Support" is calculated as a simple ratio of the number of A records to the number of total records while "confidence" is the ratio of the number of A and B records given the number of A records. In statistical terms, "support" is a frequency distribution and "confidence" is a conditional probability. Using the sample Marketing example above, a complete association rule would be written as "beer → diapers (support=2%, confidence=40%)".

More elaborate association rules are possible employing more sophisticated rule mining algorithms, such as *APriori* (e.g.,, Han and Kamber, 2001 p. 230). *APriori* is more efficient to the extent that generating *relevant* patterns does not necessarily require generating *exhaustive* (and therefore potentially computationally-infeasible) patterns. As a pilot study and an introduction to a decision-support technique that had never been done before on this campus, no individual could *ex ante*, objectively or subjectively set minimum thresholds for either the "support" or "confidence" value to be used with any more sophisticated algorithm. But we can use the support value and better, the confidence value, to *sort* the candidate rules in descending order of relevance and let the decision-makers simply use that information "as a report" in various discussions. Further, we can leverage the inherent multi-dimensional aspect of pivot tables in MS-Excel—not for a categorical bivariate analysis of unique, but exhaustive pairs (which even pivot tables can't do), but for a relatively intuitive desktop interface that permits patterns to emerge with simple inspection by each decision-maker in the scope, scale, and dimension of central interest to that decision-maker. For Chairs, that might be a review of the entire list of course patterns and interactions for each course, for Deans that might be a short list of intra-College conflicts in the core curriculum (if any), and for the University that might be a single number summarizing the average change (increase, no change, or decrease) to the number of conflicts among students taking classes in more than a single College. A multitude of other functional questions can be addressed in a similar manner.

Compared to other data mining techniques such as neural networks, association rules exhibit the least amount of statistical inference power. Recall, however, that our goal is not one of resolute generalization, but rather one of context specialization. The decision-support system just needs to "fill in the knowledge gaps" of the existing human schedulers to improve scheduling on an iterative and incremental basis. Also, by using an advanced, but mathematically simple technique, we encourage organizational decision-makers to adopt contemporary decision-support techniques to complement traditional methodologies, such as surveys and univariate descriptive techniques. Numerous resources are available that describe various data mining techniques, including association rules (e.g., Han and Kamber [2001] or Witten and Frank [1999]).

Following Gray and Watson [1998], the following data warehousing procedures were used. A source relation (table) mart was constructed by extracting attributes such as StudentID, SectionID, Course, MeetingDays, BeginTime, and NumberOfUnits from the online student information management system for the Fall, 2003 semester. Courses that were listed as "to be arranged" either by day or time (or both) were excluded. NumberOfDaysPerWeek was derived by parsing the MeetingDays attribute. The first data transformation was to estimate the EndTime

(in minutes) for each course with the following formula, [EndTime = BeginTime + (NumberOfUnits / NumberOfDaysPerWeek)*(50/60)]. The second data transformation was to expand the number of records (tuples) by breaking out the MeetingDays field (the values range from 1 to 5 days a week, but is typically two or three days per week) into a MeetingDay field, which is a single day of the week only. Other minor transformations were performed, but these are elementary in nature.

The algorithm for association rules mining is fundamentally a "join and prune" procedure. Following Han and Kamber [2001], the following "join" technique was employed. "Pruning", if needed over the longer run, can be done later. An initial working relation was constructed for each research question. As an example, the SQL statement for the functional question described in this paper is shown in Sidebar 1.

SIDEBAR I. SQL STATEMENT OF THE PROBLEM

```
SELECT DISTINCT a.FourDigitYearTerm, a.CollegeAbbr, a.DepartmentAbbr, a.Course,
a.FileNumberHash, b.FourDigitYearTerm, b.CollegeAbbr, b.DepartmentAbbr, b.Course,
b.FileNumberHash

INTO [tblExclusions--StudentCourse-2]

FROM [tblExclusions--StudentCourse-1] AS a

INNER JOIN [tblExclusions--StudentCourse-1] AS b ON (a.Course<>b.Course) AND
(a.FileNumberHash=b.FileNumberHash)

WHERE a.FourDigitYearTerm="20034";
```

The critical component of the SQL statement in Sidebar I is the unusual INNER JOIN element. This particular SQL syntax forms a "self-join", which is one type of the "equi-join" construct. The INNER JOIN element duplicates the original table internally and then compares the tuples in one table with the tuples in the other table (in the SQL statement, the source table is referred to by the alias "a" and the duplicate, internal table is referred to by the alias "b"). This SQL statement generates the initial working relation with the appropriate attribute-value pairs, effectively doubling the columns in the table. In the context of this functional question, each tuple represents a binary association between enrolled courses for each student. Slightly more complex SQL statements were instantiated for other functional questions requiring different degrees of association and fewer or greater attributes. Note that "FileNumberHash" is simply a derived field (generated randomly) which retains the uniqueness of the individual student record (so tuples can be generated correctly), but masks the identity of the individual student (so as to preserve confidentiality). Note also that "20034" refers to the fourth semester (i.e., Fall) in the year 2003.

At this point in the mining workflow, various frequency distributions were derived from the initial working relation using one or more atomic attributes of the initial working relation and summarized in various tables utilizing a conceptual hierarchy consisting of section, course, department, college, and university. Although the primary data of interest is at the nominal (categorical) level (e.g., "ACCT 220), the natural conceptual hierarchy of section-course-department-college-university is used quite well within the data mining paradigm.

The only summary calculations made during this project were the two elementary data mining measures of "interestingness" (Han and Kamber, 2001 p. 147), 'support" and "confidence" described previously. In the end, even these deterministic calculations had to explained to at least two key decision-makers in person. Athough not studied further in this particular functional

context, part of the issue seems to be that these measures are "too new" to be adopted by key decision makers without corresponding explanation. While the reward of data mining is rigorous and relevant results beyond that of descriptive statistics, the risk is that the methods will be perceived as "too complex" and therefore, unreliable or invalid. To ameliorate this situation, the pilot study focused on providing all of the course-section combination detail, rather than focusing solely on summary statistics. All reports were delivered in MS-Excel to decision-makers who reuested them. . To aid in exploratory data analysis, lists were ordered in descending order of simple frequency distribution (essentially, the data mining "support" measure, just stated as an absolute value rather than a proportion). Pivot tables for each list were also created. Pivot tables proved extremely useful in this situation, because

1. they seem to be "trusted" by several of the decision-makers, and
2. the data in a single pivot table scales to provide summary answers for decision-makers at the University-level and at the same time, detailed answers for decision-makers at the chair level.

VI. RESULTS

In this particular analysis, the source relations, the intermediate relations generated by the INNER JOIN, the initial working relations, and the tables holding the data with relevant reporting measures consumed approximately 1.6 GB in a single MS-Access 2000 database. Although large in size, this size is suitable for replication for a single semester in the future. Execution of this SQL statement takes approximately 45 minutes on a Pentium III 450 Mhz desktop PC. The SQL statement shown can be reused for other functional questions primarily by modifying the JOIN and WHERE expressions. The practical value of attempting to answer difficult functional questions with modest hardware and software technology cannot be understated. This approach provides not only an insight into the computational boundaries of the problem space, but also a perspective on how difficult replicating the data mining procedure to other problems, possibly with other analysts, will eventually be.

The exclusionary effect to students enrolled in sections in the pilot study is negligible. The hypothetical maximum number of students that would experience a conflict is approximately 12%. But by evaluating the actual enrollment patterns more closely, the actual number is slightly less than 2%. Even for the 2% of the students that are impacted, the reduction is just in the number of course-sections for a single course at a single time, and not in the variety of courses overall. Further, this 2% result is without any mitigating changes in any other College. Of the other courses in the other Colleges, one course (a MATH course) was involved in the majority of the conflicts due to it often being scheduled for one hour per day for five days per week. This negligible result is also due in part to the small increase in the "intersection rate" (or "overlap") of 1.5 hour sections aberrantly scheduled on Monday and Wednesday and partially due to (chiefly) upper-division students in this College form natural "dispersion groups" with respect to enrolling in specific courses originating in other Colleges. These findings tend to support the anecdotal observation by the academic decision-makers in the pilot College that no complaints from the approximately 1,080 (27 sections * 40 students/section) students impacted escalated a "scheduling exclusion" complaint. These findings also tend to support some preliminary, but growing evidence that students prefer "two-day-a-week, 1.5 hour courses" to "three-day-a-week, 1 hour courses." [Weiss, 2004]. Revisiting existing course scheduling policies at similar universities with similar student populations may become a requirement for some universities [Sampson, et. al., 1995].

VII. DISCUSSION

University timetabling remains an active area of research (e.g., , [Asratian and de Werra, 2002]), however, it is not clear that methodological approaches of reductionism that continue to decompose the problem into finer and finer granularity with earlier and earlier lead times before the semester starts is the most prudent strategy for some institutions. Decision-support systems,

including systems that incorporate appropriate course scheduling modules, are extremely useful in the academic process [Murray, et. al., 2000; Kassicieh, et. al., 1986]. And partially due to its complexity, course scheduling may be viewed as one of the best processes to understand well in an academic business process re-engineering (BPR) context [Denning and Median-Mora, 1995].

From an information systems perspective, many challenges must be faced to manage and use the extremely large amounts of data already captured in existing systems. One of them is the pressing need to learn about the phenomenal growth in machine learning and other heuristic techniques that arose chiefly out of the computer science discipline in the past decade. Another is the need to re-examine the boundaries between disparate system interfaces. For example, is the task of deriving the candidate association rules (e.g, the course pairs) a data "warehouse" function (more like "information technology") or a data "mining" function (more like "institutional research") or something else in between? Still another is how to educate academics and practitioners on the strengths and limitations of such decision-support methods. Finally, as Information Systems academics, we should be able to combine our natural strengths as "boundary-spanners" with our rich understanding of the quantitative histories of one or more referent disciplines. At a minimum, addressing these types of thorny applied problems in our own institutions as "consulting" or "service" work seems to be a good fit with the differential value proposition engendered by the skill sets of Information Systems academics.

The results of this data mining effort challenged core beliefs about student demand and the degree to which time-slot overlaps, especially inter-College overlaps, may have on students. Organizationally, it is one thing to challenge a deeply-held belief given a certain set of circumstances; it is quite another to advocate for operational implementation of a new paradigm. This pilot study suggests that new scheduling options can be operationalized and new methods to evaluate the consequences of such a schedule can be analyzed efficiently. One unintended result of the findings generated from this data mining analysis was to spark a formal and thorough review of existing University scheduling policies and practices. This analysis and design effort is expected to take approximately one year.

VIII.CONCLUSIONS

While approaches to the optimization problem of course scheduling performed before the semester begins may be desirable, such approaches may not be possible in some cases. Alternate approaches, including inductive data mining methods, may be the best, or in some cases, the only, analytic technique. Results of any given data mining approach are best evaluated iteratively by all stakeholders collaboratively and can be used to triangulate other results derived from more traditional data collection techniques such as surveys and focus groups. To the extent that course scheduling at the University studied will continue to be managed in a decentralized ("persistently federated") manner for the foreseeable future, interventions such as small-scale pilot studies may be the only workable strategy to provide empirical data on aberrant scheduling. The optimal approach in the long run may be an interactive approach (see, e.g., [Ferland and Fleurent, 1994]) that combines a prescribed, but flexible suite of time-slots with a process that provides iterative and incremental feedback for time-slot model improvement [Dimopoulou and Miliotis, 2004].

IX.LIMITATIONS

The data mining methods and corresponding results described in this paper suffer from a number of limitations. This study evaluated the patterns that resulted from approximately 8.5% (27/314) of the number of total course-sections, and in only a single College with approximately 16% (5,500/33,000) of the students. It is possible, although unlikely, that the aberrantly-scheduled course-sections were not typical or that students were indeed negatively impacted but either didn't recognize it or did recognize it, but consciously chose not to initiate or escalate a complaint. This study also did not perform more advanced comparisons by deriving student enrollment patterns across semesters. To the extent that a study of a single semester does not capture

temporal or longitudinal changes in student activity, a single semester methodology eschews subtle changes in student demand and explicit preferences. Therefore any conclusions drawn from the data mining inferences are incomplete. This latter issue could be addressed with the same data mining approach articulated in this paper, but the single database size limitation of MS-Access (2 GB) may preclude the use of MS-Access as the SQL engine.

Finally, frequency distribution was used as a surrogate for the more typical data mining measure of interestingness, relevance. Although the equation for relevance and its data mining cousin, confidence, is simple to compute, it may be non-trivial to explain fully to all decision-makers given that this data mining effort is the first of its kind at this institution. As this general data mining framework is widely adopted as appropriate, accurate and useful by the relevant decision-makers, including research staff and senior executives, these and other, more advanced data mining techniques can be applied.

X.FUTURE WORK

The methodology described in this paper can be extended in a number of ways. Future technical work involves ensuring that all of the SQL statements employed in this research can scale to address larger problem spaces and can port to alternate database platforms. In principle, it should also be possible to address nearly all research questions regarding course scheduling, even extremely sophisticated ones, by leveraging the theory and practice of data mining. As discussed in Section V, a number of functional questions originate from the many stakeholders involved in the course scheduling process. For example, human schedulers need to know which course combinations can never be taken together in the same semester (due to how they are scheduled). Human schedulers also need to know which courses have never been taken together in the past, or more important, are not needed in any given degree of study. To the extent that faculty preferences can be induced across semesters, data mining may assist in managing this crucial, but often intangible, component of course scheduling as well [Badri, et. al., 1998]. Similarly, to the extent that student preferences can be solicited or induced, data mining assists in estimating the parameters needed in any given model.

ACKNOWLEDGEMENTS

This paper was greatly enhanced by the contributions of two anonymous reviewers. A brief presentation on the work discussed in this paper was given at the annual California Association of Institutional Research meeting in Anaheim, CA in November, 2004.

Editor's Note: This article was fully peer reviewed. It was received on July 30, 2004 and published on September 18, 2005. It was with the author for 7 months for one revision.

REFERENCES

Asratian, A. S., and D. de Werra (2002). "A Generalized Class-Teacher Model for Some Timetabling Problems", *European Journal of Operational Research*, (143)3, pp. 531-542, December 16.

Badri, M.A. (1996). "A Two-Stage Multi-Objective Scheduling Model for Faculty-Course-Time Assignments", *European Journal of Operational Research*, (94)1 pp. 16-28, October 1.

Badri, et. al. (1998). "A Multi-Objective Course Scheduling Model: Combining Faculty Preferences for Courses and Times", *Computers and Operational Research*, (25)4, pp. 333-316, April.

Benjamin, R., and E. Levinson (1993). "A Framework for Managing IT-Enabled Change", *Sloan Management Review*, (34)4, Summer, pp. 23-33.

Boronico, J. (2000). "Quantitative Modeling and Technology Driven Departmental Course Scheduling", *Omega*, 28, pp. 327-346.

Burke, E., Jackson, K., and J. Kingston (1997). "Automated University Timetabling: The State of the Art", *The Computer Journal*, (40)9, pp. 565-,.

Burke, E., Elliman, D. G., and R.F. Weare (1993). "Automated Scheduling of University Exams", in *The Proceedings of the IEEE Colloquium on Resource Scheduling for large Scale Planning Systems*, pp. 142-148.

Chang, L. (2003) "Applying Data Mining Technology in Modeling and Predicting Admissions Yields in Higher Education", *Proceedings of the 43rd Forum of the Association of Institutional Research*, Tampa, FL, May.

Cheng, T.C. (1993). "Operations Research and Higher Education Administration", *Journal of Educational Administration*, (31)1, pp. 77-90.

Cox, J., and R. Jesse (1981). "An Application of Material Requirements Planning to Higher Education", Decision *Sciences*. 12(2), pp. 240-260, April.

Delone, W., and E. R. McLean (1992). "Information Systems Success: The Quest for the Dependent Variable", *Information Systems Research*, (3)1, pp. 60-95.

Denning, P. and R. Medina-Mora (1995). "Completing the Loops", *Interfaces*, (25) 3, pp. 42-57, May-Jun.

Dimopoulou, M. and P. Miliotis (2004). "An Automated University Course Timetabling System Developed in a Distributed Environment", *European Journal of Operational Research*, (153)1, pp. 136-147, February 16.

Dinkel, J. J., Mote, J., and M.A. Venkataramanan (1989). "An Efficient Decision Support System for Academic Course Scheduling", *Operations Research*, (37)6, pp. 853-864, Nov/Dec.

Ferland, A.J., and C. Fleurent (1994). "SAPHIRL A Decision Support Systems for Course Scheduling", *Interfaces*, (24)2, pp. 105-115, Mar-Apr.

Gray, P., and H. Watson (1998). *Decision Support in the Data Warehouse*, Englewood Cliffs, NJ:Prentice Hall.

Han, J., and M. Kamber (2001). *Data Mining: Concepts and Techniques*, San Francisco:Morgan-Kaufmann.

Hertz, A. (1991). "Tabu Search for Large Scale Timetabling Problems", *European Journal of Operational Research*, (54)1, pp. 186-193, September 5.

Hertz, A., and V. Robert (1998). "Constructing a Course Schedule by Solving a Series of Assignment Type Problems", *European Journal of Operational Research*, (108)3, pp. 585-603, August 1.

Hinkin, T. R., and G.M. Thompson (2002). "SchedulExpert: Scheduling Courses in the Cornell University School of Hotel Administration", *Interfaces*, (32)6, pp. 45-57, Nov-Dec.

Juan, L. (2002). "Using Unsupervised and Classification and Regression Trees Algorithms to Develop a Student Learning Outcomes Typology", *Proceedings of the Annual Conference of the California Association of Institutional Research*, Pasadena, CA, May.

Kassicieh, S. K., Burleson, D. K., and R.J. Lievano (1986). "Design and Implementation of a Decision Support System for Academic Scheduling", *Information and Management*, (11)2, pp. 57-64, September.

Keast, D. (1998). "Part-time University Education", *The International Journal of Educational Management*, (12)3, pp. 114-, September.

Kehoe, B. (2004). Personal Communication, Director of Institutional Research, California State University, Fresno, February.

Kiaer, L. and J. Yellon (1992). "Weighted Graphs and University Course Timetabling", *Computers and Operational Research*, (19)1, pp. 59-67, January.

Kumar, T. (2004) "Data Mining: The Next Revolution in Institutional Research", *Proceedings of the 44rd Forum of the Association of Institutional Research*, Boston, MA, May.

Mooney, E. L. Rardin, R.L., and W.J. Parmenter (1996). "Large-scale Classroom Scheduling", *IIE Transactions*, (28)5, pp. 369-378, May.

Mannila, H. (2000). "Theoretical Frameworks for Data Mining", *ACM SIGKDD Explorations*, (1) 2, January, pp. 30-32.

Murray, W. S., LeBlanc, L. A., and C. Rucks (2000). "A Decision Support System for Academic Advising", *Journal of Organizational and EndUser Computing*, (12)3, pp. 38-49, Jul-Sep.

Rogers, E. (2003), *Diffusions of Innovation* (5th ed.), New York: Free Press.

Sampson, E. R., Freeland, J. R., and E. N. Weiss (1995). "Class Scheduling to Maximize Participant Satisfaction", *Interfaces*, (25)3, pp. 30-41, May-Jun.

Stallaert, J. (1997). "Automating Timetabling Improves Course Scheduling at UCLA", *Interfaces*, (27)4, pp. 67-81, Jul/Aug.

SPSS. (2004). Statistical Program for the Social Sciences, 2004 http://www.spss.com.

Tripathy, A. (1984). "School Timetabling—A Case in Large Binary Integer Programming", *Management Science*, (30)12, pp. 1473-1489, Dec.

Weiss, E. (2004), Personal communication, Chair, Department of Accounting and Information Systems, California State University, Northridge, March.

Willett, T. (2003) "Increasing Persistence with an Experimental Intervention Directed by Data Mining and Statistical Predictive Models", *Proceedings of the 43rd Forum of the Association of Institutional Research*, Tampa, FL, May.

Witten, I., and E. Frank (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco: Morgan-Kaufmann.

Yu, M., and C. Yang (1993). "A Simple Optimal Parallel Algorithm for the Minimum Coloring Problem and Interval Graphs", *Information Processing Letters*, (48)1, pp. 47-51, October 28.

ABOUT THE AUTHOR

Wayne Smith is a doctoral student in the School of Information Systems and Technology at Claremont Graduate University. Wayne has worked in higher education in the areas of systems and technologies for 19 years. Wayne's research interests are in the areas of data analytics and information visualization, semantic web technologies, telecommunications and internetworking, and strategic management.

Communications of the Association for Information Systems

ISSN: 1529-3181