Discovering Enrollment Knowledge in University Databases

Arun P. Sanjeev and **Jan M. Żytkow**†

Office of Institutional Research and Computer Science Department
Wichita State University, Wichita, KS 67260-0113; U.S.A.
†also Institute of Computer Science, Polish Academy of Sciences
sanjeev@cs.twsu.edu zytkow@wise.cs.twsu.edu

Keywords: Discovery in databases, incremental mining, university enrollment, contingency tables.

Abstract

We describe a data mining application in large databases of student records. Our goal has been to discover knowledge useful in understanding the university enrollment and to find ways to increase it. We demonstrate a combination of automated discovery with human involvement in the discovery process. Human operators formulate open questions and interpret the knowledge discovered by the automated discovery system. Some surprising discoveries we have made have led to the repeated cycle of asking questions, running the automated search, and interpreting new results. In this paper we focus on several findings. We show that good high school students are the best source of large numbers of credit hours, but that some of these students drop out, causing significant enrollment losses. We examine the effect of financial aid on retention. We demonstrate that remedial instruction does not seem to help retain the academically underprepared students. Our results have been surprisingly stable when we used the Fall '87 cohort to verify the findings obtained from the cohort of Fall '86. We have presented our findings to university administrators in a number of meetings. The discovered knowledge can affect decision making and policy formation.

Introduction

Several knowledge discovery systems (EXPLORA Klösgen, 1992; KDW: Piatetsky-Shapiro and Matheus, 1991; 49er: Żytkow & Zembowicz, 1993) have been developed and applied to large-scale exploration of databases in various domains. Findings from many applications have been reported in earlier workshops (Druzdzel and Glymour, 1994; Matheus, Piatetsky-Shapiro and McNeill, 1994; Smyth, Burl, Fayyad and Perona, 1994). However, still plenty of attention must be focussed on the knowledge discovery process before we can make generalizations.

The primary purpose of this paper is to discuss our experience of discovering knowledge in an educational domain. We present our discovery goals and the steps made to reach them. We discuss problems that we encountered and the way in which we re-focussed our discovery process to extract new and useful knowledge.

The problem of declining university enrollment: Student enrollment can be critical for universities. Our institution is experiencing enrollment decline which concerns both administrators and local community. In Kansas, resource allocation to state universities is driven by the number of hours the students enroll in classes. Therefore, a continuous decline in enrollment is a serious threat to the budget. But many specific steps to increase enrollment may not be productive because student enrollment is a complex phenomenon, especially in metropolitan institutions where the student population is diverse in age, ethnic origin and socio-economic status.

In order to analyze the enrollment we turned our attention to student databases at our university. We used 49er discovery system to search the databases for knowledge about enrollment. Our research originated from several open questions. Since a degree is a direct measure of student success, we asked how to increase the percentage of degrees received by students who are currently enrolled? Bachelor's degrees are awarded after completing approximately 120 credit hours. But those who do not complete a degree also take credit hours, so it is in the interest of the university that they take more hours. We also sought what caused differences in the number of terms (semesters) students enrolled. Before we present our discovery process, let us briefly describe the data and 49er's exploration method.

The cohort and the data: We wanted to start with a possibly homogenous, yet large group. We focussed on the cohort containing first-time, full-time freshmen with no previous college experience from the Fall 1986. This choice allowed sufficient time for the students to receive a bachelor's degree even after a number of stop-outs. Then we repeated the same analysis for the identical student sub-population selected from the Fall class of '87 to verify and check the stability of the patterns discovered for the cohort of Fall '86.

Student databases contain demographic and academic information. The academic information stored

for all students in each term enrolled is contained in a large number of attributes and records. We combined enrollment information for each student over all the enrolled terms. We identified our goal variables as: degrees received (DEGREE), total number of credit hours taken (CURRHRS) and the total number of academic terms enrolled (NTERM) by the students (Isaac, 1993).

We grouped independent variables into three categories. The first category describes students' *demographics:* age at first term, ethnicity, sex, and so forth. The second category describes *high school performance* (Lenning, 1982): high school grade point average (HSGPA), rank in the graduating class (HSRANK), and the results on standardized tests (COMPACT). The third category describes students' *university performance*: hours of remedial education in the first term, performance in basic skills classes during the first term, cumulative grade point average (CUMGPA), number of academic terms skipped, maximum number of academic terms skipped in a row, number of times changed major (Isaac, 1993), number of times placed on probation, and academic dismissal.

The automated method: 49er discovers knowledge in the form of regularities, that is statements of the form "Pattern P holds for data in range R". The examples of patterns include contingency tables, equations, and logical equivalence. Contingency tables (Bhattacharyya & Johnson, 1986) are very useful as a general tool for expressing knowledge which cannot be summarized into specialized patterns such as equations. Since personnel data lead to fuzzy knowledge, in this paper we will only consider contingency tables, although some could be approximated by equations. A range of data is a data subset distinguished by conditions imposed on one or more attributes.

49er can be used on any relational table (data matrix). It systematically searches a large number of data subsets so that it can capture many patterns that occur in limited circumstances. 49er typically finds a large number of two dimensional regularities. Initially, 49er looks for contingency tables, but if the data follow a more specific pattern, it can follow-on with a more subtle discovery mechanism, such as search in the space of equations. 49er does not count missing values, unless the user wants to see a pattern that they make. The remainder of each records in which the missing values occurs is used, so that all the data are used, in distinction to many systems which discard the entire record if it contains a missing value.

Statistical tests measure the significance and strength of every hypothesis, which is qualified as a regularity if test results exceed the acceptance thresholds for each test. Intuitively, significance means sufficient evidence. It is measured by the (low) probability that a given sample is a statistical fluctuation of random distribution. Threshold selection reflects the domain knowledge and research objectives. While in typical studies researchers accept regularities with $Q < 0.05$, 49er typically uses much lower thresholds, on the order of $Q < 10^{-5}$ because in large hypotheses spaces many random patterns look like significant regularities. 49er's principal measurement of contingency tables strength is based on Cramer's V coefficient. Both Q and V are derived from the χ^2 statistics which measures the distance between tables of actual and expected counts. For a given $M_{row} \times M_{col}$ contingency table, and a given number N of records,

$$V = \sqrt{\chi^2/(N \min(M_{row} - 1, M_{col} - 1))}.$$

V measures the predictive power of a regularity. The strongest, unique predictions are possible when for each value of one attribute there is exactly one corresponding value of the other. In those cases $V = 1$. On the other extreme, when the actual distribution is equal to expected by the attribute independence hypothesis, then $\chi^2 = 0$ and $V = 0$. V does not depend on the size of the contingency table nor on the number of records. Thus it can be used as a homogenous measure on regularities found in different subsets and for different combinations of attributes. The discovered regularities and the relevant statistical information can be viewed by the user. Inspecting each pattern, the user can decide on a further focused search for interesting regularities.

The initial discovery tasks: Our main focus was to determine *what categories of students enroll in more terms, take more credit hours and receive degrees.* Regularities were sought for all combinations of independent and goal variables. In addition, 49er sought regularities between independent variables. Some of the discoveries were so striking that later we expanded our focus to capture new phenomena, such as *drop-out behavior of academically good students* and *the effect of remedial instruction on the academically underprepared students.*

Regularities for Enrollment

49er's discovery process resulted in many regularities. In this paper, we focus mainly on a selected few, concentrating on those which were particularly surprising and called for further study. To mention a few examples of other regularities: big differences in persistence among races; students never placed on probation when compared to those placed on probation once enrolled in more terms, took more credit hours and received degrees at a higher percentage; students who changed their majors several times received degrees at the highest percentage.

Table 1-a,b shows that the age of the student negatively influences the number of the terms enrolled. This can be seen by considering negative (less than expected) and positive (more than expected) values in Table 1-b. Relatively high percentage of students who enter the university for the first time at the age of 18 enroll in more than 2 terms. That percentage decreases

Table 1: (a) Actual Counts Table (b) Differences Table for AGE vs NTERM; $\chi^2 = 81, Q = 5 \cdot 10^{-11}, V = 0.14$

NTERM	12 +	48	36	4	2	(a)
	9-11	155	86	3	6	
	6-8	143	86	3	10	
	3-5	134	90	7	4	
	1-2	227	262	43	47	
		<19	19-24	25-29	30+	AGE

NTERM	12 +	.0531	-.0029	.0341	-.5504	(b)
	9-11	.2242	-.1425	-.7208	-.5144	
	6-8	.1668	-.1141	-.7116	-.164	
	3-5	.1259	-.0453	-.307	-.6556	
	1-2	-.2259	.1280	.7279	.6423	
		<19	19-24	25-29	30+	AGE

slightly for those who entered the university at the age 19 to 24, and decreases even more for older students.

Table 1-b suggests additionally that between 2 and 3 terms and between 18 and 19 years are particularly useful split points to summarize the discovered pattern. Students under 19 years of age when compared to the older students drop-out within the first two terms at a lower percentage (32.1% vs 51.1%) and keep enrolling at a higher percentage (67.9% vs 48.9%) after their first two terms. Although difference tables are useful in noticing patterns, for lack of space we will skip them and report only the tables of actual counts, such as Table 1-a.

Table 2-a shows that the more terms have been skipped by the student, the smaller is the chance for larger number of enrolled hours. For instance, students who skipped less than 4 terms when compared to those who skipped from 4 to 7 terms take 90 or more credit hours at a much higher percentage (40.3% vs 14.5%). Table 2-b strongly indicates that the higher are the grades in high school, the better are the grades in college. One can see a fuzzy but very distinct linear relationship between average grade in college (CUMGPA) and average grade in high school (HSGPA).

Academic results in high school turned out to be the best predictor of persistence and superior performance in college. Similar conclusions have been reached by Druzdzel and Glymour (1994) through application of TETRAD (Spirtes, Glymour & Scheines, 1993). They used summary data for many universities, in which every university has been represented by one record of many attributes averaged over the student body. Since we considered records for individual students we have been able to derive further interesting conclusions.

Among the measures of high school performance and academic ability, our results indicate that high school grade point average (HSGPA) is a better predictor than either composite ACT score or the ranking in the graduating class. It was surprising to find in our data that the regularities for HSGPA offer stronger predictions than regularities based on nationally standardized ACT scores. This can be seen by comparing Tables 3-a and 3-b. Table 3-a shows a regularity which is more significant (Q : $10^{-32} vs 10^{-8}$) and also stronger (V:0.19 vs 0.15). The difference between predictions of both tables is not large, though.

Table 2: Actual Tables for TOTSKIP vs CURRHRS and HSGPA vs CUMGPA

CURR HRS	120+	285	5	0	0	(a)
	90-119	137	12	0	0	
	60-89	92	22	1	0	
	30-59	154	36	14	0	
	1-29	374	41	29	136	
	0	5	1	1	45	TOT
		< 4	4-7	8-11	12+	SKIP

$\chi^2 = 513.9, Q = 0.0, V = 0.30$

HS GPA	A	1	6	53	62	58	(b)
	B	5	17	137	55	14	
	C	38	132	295	61	9	
	D	27	100	56	11	1	
	F	0	1	1	0	0	CUM
		< 0.99	1-1.99	2-2.99	3-3.49	3.5+	GPA

$\chi^2 = 458.8, Q = 0.0, V = 0.28$

According to Table 3-a, among the students with HSGPA of 'C'/'D', the fractions of those who enroll in less than 30 hours and those who enroll in 30 or more hours are nearly the same. However, as we move to the 'A'/'B' grade categories: for each student who takes less than 30 hours, above 3 students enroll in 30 hours or more. Table 3-b, indicates a similar finding for ACT scores. Students who score above 22, enroll 3 times more frequently above 29 hours, than below 30 hours.

Tables 3-a,c,d show that analogous patterns of approximately the same strength and significance relate HSGPA with all three goal variables. Table 3-a has been discussed above. According to Table 3-d, students with a 'A'/'B' grade (HSGPA) when compared to those with a 'C'/'D' received bachelor's and associate degrees at a higher percentage (48.7% vs 19.2%). Also, the table clearly shows that the higher the HSGPA the greater the chance to receive a bachelor's or associate degree: from 0% for 'F' student to 56% for 'A' student.

New task: retention of good students

The search of student records for enrollment patterns resulted in very useful knowledge. One of the most important findings has been that high school grade point average is the best predictor of college performance. We will now look closely at the patterns for HSGPA.

Exceptions from the patterns in Tables 3-a,c,d: Student drop-out is a major issue since failure to retain the already enrolled students indicates possible failures in the system and is expensive in terms of credit hours that could be gained. From Table 3-a we know that a significant percentage of students with the highest HSGPA enroll in high number of credit hours. However, a closer inspection of Table 3-a, reveals that in the category of students with high school grade 'A'/'B', 97 students (22.9%) dropped out before completing a

Table 3: Actual Tables for HSGPA vs{CURRHRS, NTERM, DEGREE}; COMPACT vs CURRHRS

CURRHRS

120 +	0	11	102	92	73	(a)
90-119	0	13	67	26	32	
60-89	0	6	54	25	25	
30-59	0	34	100	32	22	
1-29	4	164	243	60	29	
0	0	14	17	5	3	
	F	D	C	B	A	HSGPA

$\chi^2 = 229.0, Q = 1.66 \cdot 10^{-32}, V = 0.19$

NTERM

12 +	0	10	41	26	8	(c)
9-11	0	16	107	70	42	
6-8	0	17	98	47	67	
3-5	1	42	110	31	31	
1-2	3	158	228	69	36	
	F	D	C	B	A	HSGPA

$\chi^2 = 168.4, Q = 3.14 \cdot 10^{-23}, V = 0.2$

CURRHRS

120+	78	59	108	5	(b)
90-119	44	24	38	3	
60-89	32	28	29	0	
30-59	68	43	36	0	
1-29	196	65	56	1	
0	8	4	2	0	COMP
	<19	≤ 22	≤ 29	>29	ACT

$\chi^2 = 83.13, Q = 1.91 \cdot 10^{-8}, V = 0.15$

DEGREE

Bachelor's	0	15	128	97	91	(d)
Associate	0	2	14	8	13	
No-degree	4	226	443	139	81	
	F	D	C	B	A	HSGPA

$\chi^2 = 156.78, Q = 1.50 \cdot 10^{-28}, V = 0.25$

total of 30 credit hours. Because students in this category are very likely to succeed, that is to take at least 120 credit hours, as much as 10,000 credit hours have been lost by losing these 97 students.

Table 3-c shows that 39.1% of the students who had received a 'A'/'B' grade in high school dropped out within the period of up to 5 terms. Another perspective on the same phenomenon can be noticed in Table 3-d. Among the 'A' students, a significant number (44%) did not stay to finish their degree.

Possible reasons for drop-out of good students: The percentage of the best students who drop-out is high. It is unlikely for these students to be drop-out for academic reasons. Perhaps these students transferred out to other degree-granting institutions. There is no data collected on transfer students, so we can only indirectly see some of the transfer effects. Table 3-a indicates that in the category of 'A'/'B' grade students, there were 165 students who had accumulated a minimum of 120 credit hours. But Table 3-d indicates that 188 students within the same category received bachelor's degrees. We can hypothesize that those students who took less than 120 credit hours at our university and graduated must have transferred credit hours from other institutions. Since our cohort contains only first-time freshmen with no previous college experience, these students could have only gained those credit hours during stop-outs from our university.

Another possible reason for drop out is lack of adequate financial aid. So far we did not use any information about financial aid. To be able to determine whether financial aid, if provided to these students, would help in their retention, we expanded our search to include the financial aid attributes.

New goal: does financial aid help retention? Financial aid is available in the form of grants, loans and scholarships. Eight types of financial aid was awarded to the students in each of the 8 fiscal years, yielding 64 attributes. Using them as independent variables, we looked for regularities with our goal variables.

The discovered results were surprising. No amount of financial aid seems to cause students to enroll in more terms, take more credit hours and receive degrees. For instance, the patterns reported for financial aid received in the first fiscal year represented very high probabilities of random fluctuation $Q = 0.88$ (for terms enrolled), $Q = 0.24$ (for credit hours taken) and $Q = 0.36$ (for degrees received). None would pass even the least demanding threshold of significance. These negative results stimulated us to seek regularities in the subgroups of students at two extremes of the spectrum: those needing remedial instruction and those who had received high school grade 'A'/'B'.

In the additional study of students needing remedial instruction we sought the regularities for financial aid received in the first fiscal year. The results were equally surprising since the patterns among the amount of financial aids received and the goal variables had the following high probabilities of random fluctuation: $Q = 0.11$ (for terms enrolled), $Q = 0.22$ (for credit hours taken) and $Q = 0.86$ (for degrees received). In the group of students receiving high school grade 'A'/'B', the corresponding probabilities were $Q = 0.99$ (for terms enrolled), $Q = 0.99$ (for credit hours taken) and $Q = 0.94$ (for degrees received). The above findings indicate that financial aid received by students in the first year was not helpful in their retention.

In further exploration we used the total dollar amount of aid received in every subsequent fiscal year 1988 – 1994 as independent variables. Yet again, the results were negative since all patterns could be interpreted as random with Q ranging from 0.04 to 0.99.

New task: remedial instruction

We discuss here the exploration we conducted to determine whether remedial classes help to retain students. We used REMHR (total number of remedial hours taken in the first term) as the independent variable.

The Problem: An intriguing regularity (Table 4) can be briefly summarized as: *"Students who took more remedial hours in their first term are less likely to receive*

Table 4: Actual Table for DEGREE vs REMHR (for all students)

DEGREE							
	Bachelor's	302	0	27	10	1	7
	Associate	32	0	3	3	1	0
	No-degree	735	2	119	82	10	47
		0	2	3	5	6	8

REMHR

$\chi^2 = 25.5, Q = 4.46 \cdot 10^{-7}, V = 0.136$

Table 5: Actual Tables for REMHR vs NTERM and REMHR vs DEGREE

NTERM						
12 +	7	2	1	0	2	(a)
9-11	15	6	1	0	3	
6-8	17	4	4	0	1	
3-5	31	7	8	0	4	
1-2	125	25	24	4	15	
	0	3	5	6	8	REMHR

$\chi^2 = 8.90, Q = 0.98, V = 0.09$

DEGREE						
Bachelor's	19	4	1	0	4	(b)
Associate	2	1	1	0	0	
No-degree	174	39	36	4	21	
	0	3	5	6	8	REMHR

$\chi^2 = 5.06, Q = 0.89, V = 0.1$

a degree". This is a disturbing result, since the purpose of remedial classes is to prepare students for the regular classes. For instance, students who took remedial education (REMHR $\geq$ 2) in their first term are less likely to receive bachelor's and associate degree than those who did not. The percentage of students receiving a bachelor's degree significantly decreased from 41% for REMHR=0 to 15% for REMHR=8.

Students needing remedial instruction: After brief analysis, we realized that Table 4 is misleading. Remedial instruction is intended only for the academically under-prepared students. These students experience academic difficulties and drop out at a higher rate. In order to obtain meaningful results, we had to identify students for whom remedial education had been intended and analyze the success only for those students. After discussing with several administrators, the need for remedial instruction was defined based on the following criteria: a composite ACT score of less than 20 and either having high school grade of 'C'/'D'/'F' or graduating in the bottom 30% of the class. Those students for whom the remedial instruction was intended but did not take it, played the role of the control group. 49er's results were very surprising because remedial instruction did not help the academically under-prepared students to enroll in more terms, take more credit hours and receive degrees.

For instance, Table 5-a indicates no relationship (Q = 0.98) between hours of remedial classes taken and number of terms enrolled. It means that remedial instruction does not influence the students to enroll in more terms. Table 5-b indicates that taking remedial classes does not improve the chances for a student to persist to a degree. For instance, those students who

Table 6: Fall 1987-Actual Tables for HSGPA vs CURRHRS and HSGPA vs NTERM

CURRHRS					
120 +	4	85	91	86	(a)
90-119	2	29	35	31	
60-89	9	56	35	18	
30-59	19	110	53	16	
1-29	63	289	91	28	
0	13	27	6	4	
	D	C	B	A	HSGPA

$\chi^2 = 213.42, Q = 2.58 \cdot 10^{-32}, V = 0.21$

NTERM					
12 +	2	31	20	8	(b)
9-11	7	90	70	43	
6-8	10	73	63	76	
3-5	23	100	57	24	
1-2	68	302	102	33	
	D	C	B	A	HSGPA

$\chi^2 = 147.28, Q = 2.08 \cdot 10^{-21}, V = 0.18$

Table 7: Fall 1987-Actual Tables for REMHR vs NTERM and REMHR vs DEGREE

NTERM						
12 +	8	0	2	0	0	(a)
9-11	29	5	3	0	1	
6-8	27	5	4	0	1	
3-5	36	7	8	1	2	
1-2	148	30	25	2	13	
	0	3	5	6	8	REMHR

$\chi^2 = 6.88, Q = 1.0, V = 0.06$

DEGREE						
Bachelor's	19	5	3	0	1	(a)
Associate	5	1	1	0	0	
No-degree	224	41	38	3	16	
	0	3	5	6	8	REMHR

$\chi^2 = 1.37, Q = 1.0, V = 0.04$

did not take any remedial classes, but needed them according to our criteria, received bachelor's and associate degrees at about the same percentage (10.8% vs 9.9%) when compared to those who took from 3 to as much as 8 hours of remedial class.

Verification in another cohort

We verified the patterns discovered for students starting in Fall '86, on records of students starting in Fall '87. The cohorts of 1986 and 1987 contained 1,404 and 1,307 students respectively. We used the same discovery process on the 1987 cohort and the results obtained are strikingly similar. Let us consider few examples.

Enrollment patterns: Table 6-a for students in the Fall '87 cohort corresponds to Table 3-a. The proportion of students taking over 30 credit hours compared to those taking less than 30 hours increases 3 times as we move from HSGPA 'C'/'D' to 'A'/'B', in striking similarity to the corresponding finding in Table 3-a. We can further notice that the patterns for Fall '87 vs Fall '86 are comparable in strength (V: 0.21 vs 0.19) and significance (Q: $10^{-32} vs 10^{-32}$). Similarly, in Table 6-b and Table 3-c: the proportion of students enrolled in more than 2 terms when compared to those

who stayed fewer than 3 terms increased triple fold as we move from HSGPA 'C'/'D' to 'A'/'B'. The patterns are also comparable in strength (V: 0.18 vs 0.20) and significance (Q: $10^{-21} vs 10^{-23}$).

Remedial courses: Table 5-a indicates no relationship (probability of random fluctuation Q=0.98) between hours of remedial classes and number of terms enrolled for students in the Fall '86 cohort. Similarly for the Fall '87 cohort, Table 7-a indicates no relationship (Q=1.0) between the same attributes. Also there is no relationship between remedial hours taken and degrees received (Fall '86: Q=0.88, Table 5-b; Fall '87: Q=1.0, Table 7-b). The consistency in the patterns discovered for the two cohorts emphasizes the severity of problems identified in this paper.

Open questions

The findings discovered by 49er have opened the door to several new questions and to many possible answers that require further exploration. We discuss briefly some of the new possibilities we will try in the future.

Enrollment: We determined that high school grade is a better predictor than ACT scores for our goal variables. It will be interesting to investigate regularities between HSGPA and ACT scores. Also, financial aid received in any one year did not help in retention. It should be determined if total aid received in all the years and other summaries of financial aid would make a difference. If so, we should then determine the preferred number of years in which the dollar amount can be distributed. Attempts should be made to find if grants/scholarships are preferred to loans and also how financial aid attributes can be combined to detect stronger patterns in the data. Also, we must use 49er to seek conditional and joint dependencies that may exist in the data to refine our conclusions.

Remedial courses: Despite selecting various subgroups of students and using many combinations of attributes, the results indicate that remedial classes did not help to retain the academically under-prepared students. This result proved to be very concerning when presented at a conference on institutional research and also at the deans council of our university. Many possibilities have been raised. Perhaps remedial programs work for other institutions. Perhaps the remedial programs should be evaluated by other criteria? Should we redefine the students who need remedial instruction? Should we investigate whether remedial instruction helps students in their subsequent basic skills classes? Perhaps the existing remedial classes should be revised? We must emphasize that none of these questions undermine our results. While the concerns are worthwhile, the usefulness of remedial instruction cannot be alleged, but must be empirically demonstrated by alternative studies.

Conclusions

We have discussed the process of discovering knowledge about student enrollment. The data were obtained from the large student databases and explored by 49er. We started with several open questions on student enrollment (who enroll in more terms, take more credit hours and receive degrees). The automated exploration conducted by 49er resulted in many interesting findings and surprises, motivating us to expand our exploration and leading to new findings and questions.

In this paper we described several particularly striking findings. We have shown that good high school students are the best source of large numbers of credit hours, but that some of those students drop out, causing significant enrollment losses. We have demonstrated that remedial instruction does not help to retain the academically under-prepared students and financial aid fails to help in retention. We have been surprised by the stability of regularities when we used the Fall '87 cohort to verify the findings obtained from the cohort of Fall '86.

References

Bhattacharyya, G.K., and Johnson, R.A. 1986. *Statistical Concepts and Methods.* Wiley: New York.

Druzdzel, M., and Glymour, C. 1994. Application of the *TETRAD II* Program to the Study of Student Retention in U.S. Colleges. In Proc. of the AAAI-94 KDD Workshop, 419-430.

Isaac, D. P. 1993. Measuring Graduate Student Retention. *New Directions for Institutional Research* 80:13-25: Baird, L.L. ed.

Klösgen, W. 1992. Patterns for Knowledge Discovery in Databases. In Proc. of the ML-92 Workshop on Machine Discovery, 1-10. National Institute for Aviation Research, Wichita, KS: Żytkow J. ed.

Lenning, T.O. 1982. Variable-Selection and Measurement Concerns. *New Directions for Institutional Research*:35-53. Pascarella, T.E. ed.

Matheus, J.; Piatetsky-Shapiro, G.; and McNeill, D. 1994. An Application of KEFIR to the Analysis of Healthcare Information. In Proc. of AAAI-94 KDD Workshop, 441-452.

Piatetsky-Shapiro, G. and Matheus, C. 1991. Knowledge Discovery Workbench. In Proc. of AAAI-91 KDD Workshop, 11–24. Piatetsky-Shapiro, G. ed.

Smith, P.; Burl, M.; Fayyad, U.; and Perona, P. 1994. Knowledge Discovery in Large Image Databases: Dealing with Uncertainties in Ground Truth. In Proc. of AAAI-94 KDD, 109-120.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causality, Statistics and Search.*

Stevens, J. 1986. *Applied Multivariate Statistics for the Social Sciences,* Lawrence Earlbaum, Hillsdale, NJ.

Żytkow, J.; and Zembowicz, R. 1993. Database Exploration in Search of Regularities. Journal of Intelligent Information Systems 2:39-81.