DOCUMENT RESUME

ED 387 017 HE 028 598

AUTHOR Sanjeev, Arun P.; Zytkow, Jan M.
TITLE Automated Knowledge Discovery in Institutional Data To Support Enrollment Management. AIR 1995 Annual Forum Paper.
PUB DATE May 95
NOTE 24p.; Paper presented at the Annual Forum of the Association for Institutional Research (35th, Boston, MA, May 28-31, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Persistence; College Credits; College Graduates; *College Students; Grade Point Average; Higher Education; *Institutional Research; Predictor Variables; *Remedial Instruction; *Student Attrition; *Student Financial Aid
IDENTIFIERS *AIR Forum

ABSTRACT

Enrollment behavior of university students was studied through a search for regularities in the enrollment data using the automated discovery system Forty-Niner (49er). Information is provided on the 49er system, the meaning of regularity, and equations used in the analysis of enrollment behaviors. Cohorts of first-time, full-time freshmen beginning college in fall 1986 (n=1,404) and fall 1987 (n=1,307) were studied with attention to degrees received, total number of credit hours taken, total number of academic terms enrolled, student demographic characteristics, and high school academic variables. It was found that high school grade point average was the best predictor of college performance and persistence. Additional findings included: within six academic terms, 39.1 percent of the students who had received an A/B grade in high school dropped out of the university; financial aid helped with student retention only when considered cumulatively over a number of fiscal years; larger loans seemed to produce more credit hours than small grants; and remedial instruction did not help students who were academically underprepared to enroll in more terms, take more credit hours, or receive degrees. (Contains 18 references.) (SW)

Automated Knowledge Discovery in Institutional Data to Support Enrollment Management

Arun P. Sanjeev
Acting Director, Office of Institutional Research
Wichita State University, Wichita, KS 67260-0113; U.S.A.
email: sanjeev@cs.twsu.edu
phone: (316)-689-3015

Jan M. Żytkow ‡
Professor, Computer Science Department
Wichita State University, Wichita, KS 67260-0083; U.S.A.
‡also Institute of Computer Science, Polish Academy of Sciences
email: zytkow@wise.cs.twsu.edu
phone: (316)-689-3925

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Fifth Annual Forum of the Association for Institutional Research held at the Boston Sheraton Hotel & Towers, Boston, Massacusetts, May 28-31, 1995. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Editor
AIR Forum Publications

Abstract

In this paper, we describe a data mining application by Forty-Niner (49er), an automated discovery system. We used student records from existing databases. Our goal has been to discover regularities in the data that can be used to understand university enrollment. We show that good high school students are the best source of large numbers of credit hours, but some of these students drop out, causing significant enrollment losses. We found that financial aid helps only when considered cumulatively over a number of fiscal years. Larger loans may produce more credit hours than small grants. We demonstrate that remedial instruction does not seem to retain the academically under-prepared students. We use a second cohort to verify our findings.

INTRODUCTION

Higher 'ucation is generally available to anyone who wishes to enter the post-secondary educational stream. The current student-body ranges widely and includes both high school graduates entering college and inquisitive senior citizens returning to sample the system again or to make a new start. Institutions. inspite of reaching out to such a broad student body, are facing difficulties in retaining the students. The enrollment problem persists and grows more severe. ,Fortunately. enrollment, persistence. and drop-out behavior of students have been stu .ied extensively (Tinto, 1975; Pascarella and Terenzini, 1980; Metzner. 1989; Krotseng, 1992; Stage, 1993). Many problems and hypotheses about causes of enrollment decline have been identified. The academic and social integration of a student into an individual institution influences persistence in higher education. Various studies indicate a variety of variables that can measure academic and social adjustments of the students. College performance affects persistence (Tinto, 1975; Krotseng, 1992). Low commitment to the institution and to the goal of degree completion often leads students to transfer (Tinto, 1975). Background characteristics like age, gender, ethnicity, and the parents' education influence persistence. The mother's education has been shown to have both postive and negative effects on persistence, depending on the students' purpose of higher educat.on. For instance, students with highly educated mothers persist in college if their goal is to improve knowledge. However, the same students attrite if their purpose is to attain a degree or to get a job (Stage, 1989). High quality advising improves college performance. which in turn decreases drop-out behavior in freshmen (Metzner, 1989). Also student-faculty informal contacts increase persistence in college (Pascarella and Terenzini, 1980).

A majority of the previous studies employed a questionnaire or a survey to collect the data for analysis. Such methods of data collection can capture data on commitments, goals, attitudes, beliefs. and the like. which have been shown to relate to persistence. However, these methods are time consuming, expensive. difficult to validate and yield fewer number of cases (records) owing to low response rates. Moreover, the study of persistence was usually not extended beyond the returning year of education (Pascarella and Terenzini. 1980; Metzner 1989; Krotseng, 1992; Stage. 1993). Any alternative that can eliminate the difficulties involved in data collection through surveys, but can still provide us with sufficient knowledge to understand persistence will be useful. Our main purpose is to gain knowledge about enrollment at our university, based on empirical generalizations of data. Students in urban institutions like ours are less likely to enroll continuously Therefore. an important component of our research was to develop a convenient and brief summary of

enrollment history for each student. We wanted to see which claims made by previous studies are confirmed by our data and whether we can find new interesting regularities about enrollment.

We studied the enrollment behavior of students through an automated search for regularities in the enrollment data by a discovery system Forty-Niner (49er). Automated exploration of databases for knowledge of various types is a well-established discipline. Several knowledge discovery systems (EXPLORA Klösgen, 1992; KDW: Piatetsky-Shapiro and Matheus, 1991; KEFIR: Piatetsky-Shapiro and Matheus, 1994; 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 several knowledge discovery workshops (Druzdzel and Glymour, 1994; Matheus, Piatetsky-Shapiro and McNeill, 1994; Smyth, Burl, Fayyad and Perona, 1994; Sanjeev and Żytkow, 1995a, 1995b). We present our goals for discovering knowledge about enrollment and the steps made to reach them. Also, we discuss problems that we encountered and the way in which we re-focussed our discovery process to extract patterns that provide new and useful knowledge.

The problem: declining university enrollment

Student enrollment can be critical for universities. Our institution is experiencing enrollment decline which concerns both administrators and the 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. Our research originated from several open questions: Degree is a direct measure of student success. Therefore, 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 degrees also take credit hours, and it is the interest of the university that they take more hours rather than less. To find out the contributing factors we also sought 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 enrolled in each term yields a large number of variables 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 ACT 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, current cumulative grade point average, 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.

A large number of patterns potentially exist in the data for the above independent variables. Also, the existing patterns differ widely in their types and may be present in all and subsets of data. The variety of subsets make the hypotheses space very large, reaching billions of hypotheses even for moderate size databases. We need a discovery system that can automatically search the large space of hypotheses, consider large number of hypotheses, discover regularities, and report the significant patterns after evaluation. The automated system must provide the user with the capability to control the automated search for regularities. For example, the user should be able to select the type of regularity to be discovered and also specify the thresholds for evaluating the discovered regularities. This is necessary because data in databases is sparse and equation-like regularities may not exist in the data. 49er is one such discovery system and it can be used on any relational table (data matrix).

49ER'S AUTOMATED METHOD

There has been considerable interest in recent years in automated methods of mining databases for useful knowledge. Data mining is attractive for many reasons. Large databases are common in all types of enterprises including educational institutions and they are uniform in structure. Plenty of useful knowledge is present in the databases, yet it is implicit. They were never systematically explored with open questions conducive to discovery. The automation of search for regularities is necessary because the forthcoming databases will encompass gigabytes or even terabytes of information.

What is a Regularity?

49er concentrates primarily on knowledge in the form of regularities. A regularity holds in a domain D if some events (situations) which are *a priori* or *logically* possible, occur in D, while some others never occur. In other words, a pattern of events exists in D, and those possible events which do not belong to that pattern never occur in D. Statistical regularities can be weaker. They hold when some combinations of events are more probable than others. This is arguably the most general characteristics of a regularity.

Regularities are statements of the form "Pattern P holds for all data in range R". The examples of patterns include contingency tables, equations, and logical equivalence. Contingency tables are very useful as a general tool for expressing knowledge which cannot be summarized into specialized patterns such as equations (Bhattacharyya & Johnson, 1986). A range R of data is a data subset distinguished by conditions imposed on one or more attributes.

49er systematically searches a very large number of data subsets in which patterns may appear. 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.

Regularities considered by 49er

49er discovers many types of regularities: 2×2 Contingency tables (in which the values of each variable are collapsed into two groups), $M \times N$ Contingency tables (in which all the values of each variable are shown) and equations. A 2×2 Contingency table is a simple, brief summary, similar to the correlation coefficient, while a $M \times N$ Contingency table shows the dependency in detail. Equation-like dependencies usually do not exist in institutional data. Therefore, we focused on $M \times N$ contingency table regularities in the data and

Table 1: Actual Counts Table for MAJCHANGE vs DEGREE

DEGREE								
	Bachelor's	41	106	65	32	12	3	0
	Associate	5	7	7	8	1	1	0
	No-degree	419	165	54	19	8	1	1
		0	1	2	3	4	5	7

MAJCHANGE

have included only those tables in this paper. As an example, we describe in detail the $M \times N$ contingency tables discovered by 49er for a regularity in the data.

Consider a regularity between the number of times students changed majors (MAJCHANGE) and received degrees (DEGREE). Table 1. shows a $M \times N$ contingency table (also called actual counts table) for those variables. Each entry in the table is equal to the number of students having the corresponding combinations of values of both variables. For example, there are 106 students who changed major once (MAJCHANGE=1) and received bachelor's degrees.

The actual counts table must be compared with the expected counts table, which is based on the null hypotheses of variable independence. 49er uses frequency distributions of the variables which form the table, to estimate the expected counts. Let $h(x)$ be the histogram of the variable x,

$$h(x) = \{(x_1, n_{x_1}), (x_2, n_{x_2}), \ldots, (x_k, n_{x_k})\},$$

where $x_1, \ldots, x_k$ are values of x while $n_{x_1} \ldots, n_{x_k}$ are numbers of records with corresponding values of x.

This way, the one dimensional knowledge derived from each variable is subtracted from the two dimensional pattern. The pattern can be visualized and its significance and strength determined. In our example:

$$h(\text{MAJCHANGE}) = \{(0, 465), (1, 278), (2, 126), (3, 59), (4, 21), (5, 5), (7, 1)\},$$

$$h(\text{DEGREE}) = \{(No-degree, 667), (Associate, 29), (Bachelor's, 259)\}.$$

If attributes x and y are independent, their joint distribution should be close to the product of distributions (histograms) of each variable. The expected number of records with $x = x_i$ and $y = y_j$ is

$$E_{ij} = \frac{n_{x_i} \cdot n_{y_j}}{N},$$

where N is the total number of records. E_{ij} is usually called the expected counts (see Table 2). Table 3 exemplifies relative differences between actual and expected counts,

$$\delta_{ij} = \frac{A_{ij} - E_{ij}}{E_{ij}}.$$

Table 2: Expected Counts Table for MAJCHANGE vs DEGREE

DEGREE								
	Bachelor's	126.11	75.40	34.17	16.0	5.70	1.36	.27
	Associate	14.12	8.44	3.83	1.80	0.64	0.15	.03
	No-degree	324.77	194.16	88.0	41.21	14.67	3.49	.70
		0	1	2	3	4	5	7

MAJCHANGE

Table 3: Differences Table for MAJCHANGE vs DEGREE

DEGREE								
	Bachelor's	-.67	.41	.90	.99	1.11	1.21	-1
	Associate	-.65	-.17	.83	3.47	0.57	5.59	-1
	No-degree	.29	-.15	-.39	-.54	-.45	-.71	.43
		0	1	2	3	4	5	7

MAJCHANGE

$\chi^2 = 219.94, Q = 4.5 \cdot 10^{-39}, V = 0.34$

49er computes all three tables. Actual counts table allows the user to make predictions of values of one variable based on the values of other variables. Expected counts table provides frequencies expected under the assumption of variable independence, while the differences table reveals unexpected events (records) in data, and visualizes details of dependency pattern between variables.

Contingency tables are a useful tool for pattern visualization, engaging the user's pattern recognition mechanisms and allowing the user to decide on the most useful description. For example, the regularity in Table 3 shows positive and negative values, which represent greater than expected and less than expected values in the corresponding cells in the actual counts table. The values .90, .99, 1.11 and 1.21 for MAJCHANGE=2 to 5 respectively indicate that students changing majors from 2 to 5 times receive bachelor's degrees more than expectedly. The '-1s' indicate no student changed majors 7 times and received a degree. Also, students who never changed majors (MAJCHANGE=0) when compared to those who changed majors from 1 to 3 times (MAJCHANGE=1 to 3) , received bachelor's degrees at a lower percentage (8.8% vs 43.8%).

Evaluation of Regularities

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. Threshold selection reflects the domain knowledge and research objectives. Significance is the probability Q that a given contingency table could have been generated randomly for two independent variables. While in typical studies researchers accept regularities with $Q < 0.05$, 49er must use much lower thresholds in the order of $Q < 10^{-5}$ because in large hypotheses spaces many random patterns look like quite significant regularities. 49er's principal measurement of contingency table 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 variable 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 homogeneous measure on regularities found in different subsets and for different combinations of variables. 49er provides the discovered regularities and the relevant statistical information for the visual inspection by the user. Inspecting each pattern, the user can decide whether a further focused search for interesting regularities is useful or not.

Predictive role of regularities

A contingency table can be used to reason about the domain, for instance, to make predictions. Suppose that we select a student who has changed majors three times (MAJCHANGE=3) and we want to predict whether the student receives a degree. From Table 1 we can infer that the student will receive a bachelor's degree with probability 0.6 (32/59), an associate degree with probability 0.1, while no degree with probability 0.3. Similar predictions can be made for any other value of MAJCHANGE or for any weighted combination of MAJCHANGE values.

The predictions are the weakest when the probabilities in a contingency table are equal to the values expected based on the independence of variables. Still, predictions are possible, but they are equal to those based on the frequency distributions. We hesitate to call such a distribution a regularity, but it satisfies our definition, because a random world (a combination of independent variables) holds a distinct, empirically verifiable pattern, and makes various distributions of events statistically improbable. Random patterns, however, are of little practical interest, so we will seek only regularities in which actual frequencies significantly differ from those expected apriori.

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.

Table 4: (a) Actual Counts Table (b) Differences Table for AGE vs NTERM; $\chi^2 = 81, Q = 5 \cdot 10^{-11}, V = 0.14$

(a)

NTERM	12 +	48	36	4	2
	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

(b)

NTERM	12 +	.0531	-.0029	.0341	-.5504
	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

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*, *effect of financial aid on retention* and *effect of remedial instruction on the academically under-prepared 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. Few examples of other findings: 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.

Table 4-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 4-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 slightly for students who entered the university at the age 19 to 24, and decreases even more for older students.

Table 4-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 brevity of space we will skip all further difference tables.

Table 5-a shows that the more the terms 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 5-b strongly indicates that the higher the grades in high school, the better are the grades in college.

Table 5: 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

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

Table 6: 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

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

CURR HRS	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	
		<19	≤ 22	≤ 29	>29	COMPACT

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

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

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

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 of many universities, in which every university has been represented by one record of many attributes averaged over the student body. In our study we consider records for individual students and therefore, 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 of persistence than either composite ACT score or the ranking in the graduating class. It was surprising to find that the regularities for HSGPA offer stronger predictions than regularities based on nationally standardized ACT scores. This can be seen by comparing Table 6-a and Table 6-b.

Table 6-a when compared to Table 6-b 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. According to Table 6-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 hours or more are nearly the same. However, as we move to the 'A'/'B' grade categories: for each student that takes less than 30 hours, above 3 students enroll in 30 hours or more Table 6-b, indicates a similar finding for ACT scores. Students who score above 22, enroll 3 times more frequently for 30 or more hours, than for less than 30 hours.

Tables 6-a,c,d show that analogous patterns of approximately the same strength and significance relate HSGPA with all three goal variables. Table 6-a has been discussed above. According to Table 6-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' grade student to 56% for 'A' student.

NEW TASK: RETENTION OF GOOD STUDENTS

Regularities discovered for enrollment indicated that high school grade point average influenced mostly the students performance and persistence in college. But, the discovered patterns also highlighted a problem which is very concerning. We shall now look closely at the patterns for HSGPA.

Exceptions from the patterns in Tables 6-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 lost and degrees not received. From Table 6-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 6-a, reveals that in the category of students with high school grade 'A'/'B', 97 students (22.9%) dropped out before completing a 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 6-c shows a similar phenomenon. Within the period of up to 6 terms 39.1% of the students who had received a 'A'/'B' grade in high school dropped out from our university. Another perspective on the same phenomenon can be noticed in Table 6-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 best students dropping-out is high and it is unlikely for these students to be dismissed for academic reasons. Perhaps these students transferred to other degree-granting institutions. We can only indirectly see some of the transfer effects since there is no data collected on transfer students. Table 6-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 6-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 contained 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 study by including many financial aid attributes.

NEW GOAL: DOES FINANCIAL AID HELP RETENTION?

Financial aid is available in the form of grants, loans, scholarships and work-study. Eight types of financial aid was awarded to the students in each of the 8 fiscal years (1987-1994), yielding 64 attributes. Using them as independent variables, we looked for regularities with our goal variables. Initially, we focussed on the total dollar amount of financial aid awarded to students in the first fiscal year and also in each of the subsequent fiscal years. Later, we expanded our search and looked for regularities in the data for our goal variables with financial aid received in all of the fiscal years (GTOTAL) and also looked for differences in the types of financial aid viz: loans, grants and work-study. As an example, let us consider a student who may have received a total of $4,000 as financial aid in four of the eight fiscal years. The above student may have been awarded $2,000 as grants, $1,500 as loans and the remaining $500 as work-study in the span of four fiscal years. We have sought regularities for our goal variables with the total dollar amount of financial aid received in all of the four fiscal years by the student ($4,000) and also for the financial aid received in the form of grants ($2,000), loans ($1,500) and work-study ($500). The results we discovered during our initial consideration of financial aid (first and each subsequent fiscal year) were surprising.

Financial aid awarded in individual fiscal years

No amount of financial aid awarded in a single fiscal year seemed 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 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 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). Whenever, the probability of random fluctuation does not have a value in the order of $Q < 10^{-5}$, it can be inferred that there is no regularity present in the data.

In further exploration, we used the total dollar amount of aid received in every subsequent fiscal year 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. The above results indicate that financial aid awarded in an individual fiscal year did not help in the retention of students. This finding encouraged us to seek regularities in the data for summaries of financial aid awarded to the students. We added cumulatively the financial aid awarded to students during the entire length of study. So, if a student had received financial aid in four fiscal years, then the total aid received by the student in all fiscal years would constitute only the contribution from four and none from the remaining fiscal years.

Financial aid awarded in all fiscal years

In this step, we discovered significant patterns present in the data which imply that financial aid awarded in all fiscal years helped the students to enroll in more terms, credit hours and receive degrees.

Table 7-a, represents the regularity between total financial aid received in all the years and number of terms enrolled. Students who received financial aid upto $10,408 when compared to those who received

Table 7: Actual Tables for GTOTAL vs{NTERM, CURRHRS, DEGREE}

NTERM

12 +	43	13	5	3	0
9-11	125	37	17	4	1
6-8	142	35	4	0	0
3-5	127	10	0	0	0
1-2	313	0	0	0	0
	≤10,408	≤20,817	≤31,226	≤41,636	≥41,636

(a) GTOTAL $\chi^2 = 155.1, Q = 9.7 \cdot 10^{-21}, V = 0.19$

CURRHRS

120 +	144	43	21	3	1
90-119	88	23	3	4	0
60-89	62	19	1	0	0
30-59	116	10	0	0	0
1-29	310	0	0	0	0
0	27	0	0	0	0
	≤10,408	≤20,817	≤31,226	≤41,636	≥41,636

(b) GTOTAL $\chi^2 = 163, Q = 3.2 \cdot 10^{-20}, V = 0.19$

DEGREE

Bachelor	189	49	14	3	1
Associate	28	2	3	0	0
No-degree	535	44	9	4	0
	≤10,408	≤20,817	≤31,226	≤41,636	≥41,636

(c) GTOTAL $\chi^2 = 46.2, Q = 1.3 \cdot 10^{-6}, V = 0.16$

between $10,408 and $20,817, enrolled at a higher percentage (58.6% vs 10.5%) upto 5 terms and at a lower percentage (41.3% vs 89.5%) beyond 5 terms. Table 7-b suggests a similar finding in the same categories of financial aid received by the students. Students who received less financial aid enrolled in classes at a higher percentage (60.6% vs 10.5%) upto 59 credit hours but continued taking classes at a lower percentage (39.4% vs 89.5%) beyond 59 hours. Table 7-c indicates that students who were awarded more financial aid when compared to those who were awarded less financial aid received bachelor's degrees at a higher percentage. However, there is a limit on the dollars that can be awarded as financial aid, in order to increase the number of degrees awarded. Students who received financial aid upto $10,408 when compared to those who received between $10,408 and $20,817 received bachelor's degrees at a lower percentage (25.1% vs 51.6%). The regularities discovered indicate that total financial aid awarded in all fiscal years helped to retain the students. We then aspired to see if the students preferred grants/scholarships to loans and work-study.

Therefore, we sought regularities between our goal variables and total aid received (in all fiscal years) in the form of loans (TLOANS), grants (TGRANTS) and work-study (TWKST). Financial-aid in the form of work-study did not seem to influence the students to enroll in more terms, take more credit hours and receive degrees, while grants and loans helped to retain the students.

Table 8-a shows the regularity for total aid received as grants and credit hours taken. We notice in the table that the larger the dollars awarded as grants the higher are the credit hours taken by the students. Students who received as much as $5,904 in the form of grants when compared to those who received between

Table 8: Actual Tables for (a) TGRANTS vs CURRHRS (b) TLOANS vs CURRHRS

CURRHRS							
	120 +	130	41	18	2	2	(a)
	90-119	86	14	5	0	0	
	60-89	56	14	4	0	0	
	30-59	87	13	0	0	0	
	1-29	213	1	0	0	0	
	0	17	0	0	0	0	
		≤5,904	≤11,819	≤17,714	≤23,619	≥23,620	TGRANTS $\chi^2 = 96.3, Q = 7.0 \cdot 10^{-9}, V = 0.16$

CURRHRS							
	120 +	86	31	7	0	1	(b)
	90-119	54	16	4	2	0	
	60-89	40	10	1	0	0	
	30-59	86	4	0	0	0	
	1-29	219	0	0	0	0	
	0	15	0	0	0	0	
		≤7,790	≤15,581	≤23,372	≤31,163	≥31,164	TLOANS $\chi^2 = 111, Q = 3.2 \cdot 10^{-11}, V = 0.19$

$5,904 and $17,714 enrolled in classes at a higher percentage (53.8% vs 12.7%) upto 59 credit hours but continued taking classes at a lower percentage (46.2% vs 87.3%) beyond 59 hours. Table 8-b suggests a similar finding for aid awarded as loans and credit hours taken. Students who received as much as $7,790 in the form of loans when compared to those who received between $7,790 and $23,372 enrolled in classes at a higher percentage (64% vs 5.5%) upto 59 credit hours but continued taking classes at a lower percentage (36% vs 94.5%) beyond 59 hours. Fortunately, 49er's evaluation of regularities allows us to make comparisons. The regularity discovered for loans when compared to that discovered for grants is stronger (V: 0.19 vs 0.16) and more significant (Q: $10^{-11} vs 10^{-9}$). Perhaps, these tables indicate that loans in larger amounts produce more credit hours from the students than grants that are awarded in relatively smaller dollar amounts. Similar regularities were also discovered for number of terms enrolled. The probability of random fluctuation Q and Cramer's V for NTERM with loans and grants are: Q=10^{-11}, V=0.19; and Q=10^{-9}, V=0.16 respectively.

NEW TASK: USEFULNESS OF REMEDIAL INSTRUCTION

In this section we discuss our expanded exploration of student records to determine whether remedial classes help to retain the students. We used REMHR (total number of remedial hours taken in the first term) as the independent variable and sought regularities for our goal variables. A particular regularity that we discovered refined our discovery process. We describe followingly, that regularity in detail and also the way in which we resolved the problem.

The Problem

An intriguing regularity (Table 9) can be briefly summarized as: *"Students who took more remedial hours*

Table 9: 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 10: Actual Tables for (a) REMHR vs NTERM and (b) 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

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

in their first term are less likely to receive 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 ≥ 2, to as much as 8) in their first term are less likely to receive a bachelor's or 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.

New range: students intended for remedial instruction

After brief analysis, we realized that Table 9 is misleading. Not all students need to take remedial classes because remedial instruction is intended only for the academically under-prepared students. Those who require remedial instruction are less academically prepared and therefore 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. The results obtained by 49er 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 10-a indicates no relationship (probability of random fluctuation $Q = 0.98$) between hours of remedial class taken and number of terms enrolled. No regularity detected means that remedial instruction does not influence the students to enroll in more terms. Table 10-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 11: Fall 1987-Actual Tables for HSGPA vs CURRHRS and HSGPA vs NTERM

CURR HRS	D	C	B	A	
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	D	C	B	A	
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

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

did not take any remedial class, but needed them according to our criteria, received either a bachelor's or an associate degree at more-or-less the same percentage (10.8% vs 9.9%) as those who took from 3 to as much as 8 hours of remedial class.

VERIFICATION IN ANOTHER COHORT

We decided to verify the patterns discovered for students starting in Fall '86, on records of students starting in Fall '87. The cohort of 1986 included 1404 students, while the cohort of 1987 included 1,307 students. We used the same discovery process. The results obtained from the two cohorts are strikingly similar. Let us consider few examples.

Enrollment patterns

Table 11-a for students in the Fall '87 cohort corresponds to Table 6-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'. This is strikingly similar to the corresponding finding in Table 6-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 11-b and Table 6-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 10-a indicates no relationship (probability of random fluctuation Q=0.98) between hours of remedial classes taken and number of terms enrolled for students in the Fall '86 cohort. Similarly for the students in the Fall'87 cohort. Table 12-a indicates no relationship (Q=1.0) between the same attributes. Also there

Table 12: Fall 1987-Actual Tables for REMHR vs NTERM and REMHR vs DEGREE

NTERM	0	1	2	3	4	5	(a)
12 +	8	0	2	0	0		
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	0	1	2	3	4	5	(a)
Bachelor's	19	5	3	0	1		
Associate	5	1	1	0	0		
No-degree	224	41	38	3	16		
	0	3	5	6	8	REMHR	

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

is no relationship between remedial hours taken and degrees received (Fall '86: Q=0.89, Table 10-b; Fall '87: Q=1.0, Table 12-b). The consistency of patterns discovered in both the cohorts proves its stability and emphasizes the seriousness of all the problems we have 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. In this section, 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 find through a regularity between HSGPA and ACT scores, whether good students in high school score lower in ACT tests. Also, financial aid when considered cumulatively for all fiscal years helped to retain the students. We must now determine: In how many fiscal years can the dollar amount be distributed? Is there a minimum dollar amount that when awarded as aid in an individual fiscal year, help in retention? Would the type of financial aid make a difference in an individual fiscal year? Finally, we must compare our results for financial aid with the results obtanied for students receiving no financial aid.

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 when presented at the MIDAIR regional conference earlier in St. Louis and also at the deans council of our university proved to be very concerning. The reason may be because remedial classes have been shown to be effective at most colleges in the nation. Should we redefine the students who need remedial instruction? Since exit exams are provided as a substitute for remedial classes, should we examine who qualifies to take

the test? Should we take into account all hours of remedial classes taken during the entire college experience? Should we investigate whether remedial instruction helps students in their subsequent basic skills classes? or perhaps the existing remedial classes should be reviewed, revised, revoked and new classes be introduced. Answers to the above questions may provide additional knowledge, useful to determine the actions the university should take in order to retain the academically under-prepared students.

CONCLUSIONS

We have discussed the process of discovering knowledge about student enrollment. We started with several open questions on ways to increase the percentage of students who enroll in more terms, take more credit hours and receive degrees. As we pursued our analysis, more questions were raised and new goal variables were identified. The entire data was obtained from the existing large student databases at our university. We used 49er for the automated exploration and searched for regularities in the data. Our discovery resulted in many interesting findings and surprises. They, in turn encouraged us to expand our exploration leading to new findings and more questions.

In this paper we have described particularly striking findings. We confirm in our data, several findings from earlier studies. Some examples are: high school performance improves college performance and is a good predictor of persistence; age of student affects persistence. 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 big enrollment loss. We have examined the effect of financial aid on retention. We have found that financial aid received by students in the first fiscal year or any subsequent fiscal year failed to help in the retention. On the contrary, financial aid received by students when considered cumulatively from all of the fiscal years, helped in the retention of students. In addition, loans in larger amounts seem to produce more credit hours from the students than grants in relatively smaller amounts. We have also demonstrated that remedial instruction does not help the academically under-prepared students to enroll in more terms, take more credit hours and receive degrees. We used similar students starting in Fall '87 as our second cohort and followed the same discovery process. The findings obtained from the Fall '87 cohort verified and confirmed the stability of regularities discovered for students from the earlier cohort.

Acknowledgment

Special thanks to Peter Zoller and Padma Chellappan for helpful comments and suggestions.

REFERENCES

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

Druzdzel, M., & Glymour, C., 1994. Application of the TETRAD II Program to the Study of Student Retention in U.S. Colleges; in: *Proc. of AAAI-94 Knowledge Discovery in Databases Workshop*, p. 419-430.

Isaac, D. P., 1993. Measuring Graduate Student Retention, in: Baird, L. & Leonard ed. *Increasing Graduate Student Retention and Degree Attainment*, pp. 13-25.

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

Krotseng, V. M., 1992. Predicting Persistence from the Student Adaptation to College Questionnaire: Early Warning or Siren Song?, *Research in Higher Education*, Vol. 33, No. 1, pp.99-111.

Lenning, O., 1982. Variable-Selection and Measurement Concerns, in: Pascarella, T., Ernest ed. *Studying Student Attrition*, pp 35-53.

Matheus, J., Piatetsky-Shapiro, G. & McNeill, D., 1994. An Application of KEFIR to the Analysis of Healthcare Information, in: *Proc. of AAAI-94 Knowledge Discovery in Databases Workshop*, p. 441-452.

Metzner, B., 1989. Perceived Quality of Academic Advising: The Effect on Freshman Attrition, *American Educational Research Journal*, Fall 1989., Vol. 26. No. 3, pp.422-442.

Pascarella, E.,& Terenzini, P., 1980. Predicting Freshman Persistence and Voluntary Drop-out Decisions from a Theoretical Model, *Journal of Higher Education*. Vol. 51. No. 1, pp. 60-75.

Piatetsky-Shapiro, G. & Matheus, C., 1991. Knowledge Discovery Workbench, in: G. Piatetsky-Shapiro ed. *Proc. of AAAI-91 Workshop on Knowledge Discovery in Databases*, pp. 11-24

Sanjeev, A., & Żytkow, J., 1995a. Mining Student Databases for Enrollment Patterns, in: *Proc. of the ECML-95 Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases.*

Sanjeev, A., & Żytkow, J., 1995b. Discovering Enrollment Knowledge in University Databases, in: *Proc. of AAAI-95 The First International Conference on Knowledge Discovery and Data Mining (KDD-95).*

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

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

Stage, F., 1989. Motivation, Academic and Social Integration, and the Early Dropout, in: Richardson-Koehler, Virginia, (Ed.), *American Educational Research Journal*, Fall 1989., Vol. 26, No. 3, pp.385-402.

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

Tinto, V., Winter 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research, *Review of Educational Research*, Vol. 45, No. 1, pp.89-125.

Żytkow, J., & Zembowicz, R., 1993. Database Exploration in Search of Regularities, *Journal of Intelligent Information Systems*, 2, p.39-81.