Modeling the Business Process by Mining Multiple Databases

Arun P. Sanjeev and Jan M. Żytkow †

Office of Institutional Research
Wichita State Univ., Wichita, KS 67260-0113; U.S.A.
†Computer Science Department
UNC Charlotte, Charlotte, N.C. 28223, U.S.A.
also Institute of Computer Science, Polish Academy of Sciences
sanjeev@cs.twsu.edu zytkow@uncc.edu

No Institute Given

Keywords: Discovery in databases, process understanding, knowledge from multiple tables, university enrollment.

Abstract. Institutional databases can be instrumental in understanding a business process. Additional databases can be also needed to broader the empirical perspective on the investigation. We present a few data mining principles by which a business process can be analyzed in quantitative details and new process components can be postulated. Sequential and parallel process decomposition can apply, guided by the results of data mining and human understanding of the investigated process. In a repeated cycle human operators formulate open questions, use queries to get relevant data, use quests that invoke automated search, and interpret the discovered knowledge. As an example we use mining for knowledge about student enrollment, which is an essential part of the university educational process. The target of discovery has been quantitative knowledge useful in understanding the university enrollment. Many discoveries have been made. The particularly surprising findings have been presented to the university administrators and affected the institutional policies.

1 Business process analysis

Understanding the business process is critical to the successful functioning of any business. Many databases have been developed to store detailed information about business processes. By design, the data capture the key information about events that add up to the entire process. For instance, university databases keep track of student enrollment, grades, financial aid and other key information recorded each semester or each year.

Knowledge discovered in data can facilitate process understanding. In addition to the known elements of the process, which have been used in database design, further knowledge can be discovered by empirical analysis. We can use a discovery mechanism that mines a database in search of knowledge then can be used to postulate a particularly justified hidden structure within the process. For instance, it may turn our that different groups of students finish their degrees in different proportion or that they take drastically different numbers of credit hours.

In this paper we focus on knowledge automatically derived from data, in distinction to expert knowledge. We present a discovery process that results in knowledge which aids the business process understanding. The discovery process is driven by two factors. The first is the basic structure of the business process and the way it is represented by database schemas and attributes. The second factor is the data and empirical knowledge they provide. The first is used to plan data preparation and design the search for knowledge. The data are then searched and the knowledge discovered from data can lead to further data preparation and further data mining.

Quests generate knowledge while queries prepare data. Automated search for knowledge can use discovery systems such as EXPLORA Klösgen, 1992; KDW: Piatetsky-Shapiro and Matheus, 1991; 49er: Żytkow & Zembowicz, 1993; KDD-R: Ziarko & Shan, 1994; Rosetta: Ohrn, Komorowski, Skowron, Synak, 1998. A discovery system requires a well defined search problem, which we call a quest. It also requires data, which are typically a subset of all available data, and are obtained in response to a query. Queries can be supported by a DBMS, but when data come from several databases, data miners must use their own query execution programs. An extended KDD process can be described by a sequence of quests and queries.

Database design knowledge includes temporal relations between atributes. When data describe a business process, the temporal order of events captured by different records and attributes is clear most of the time. Since an effect cannot precede the cause, questions about possible causal relations are constrained within the temporal relation between attributes. The sophisticated search for causes, performed by systems such as TETRAD (Spirtes, Glymour & Scheines, 1993) can be substantially reduced. A typical question about causes that still can be asked is "which among the temporarily prior attributes influences a given set of target attributes?"

Vital characteristics of a process include throughput, output and duration. No doubt that the output and/or throughput are the most important effects of each business process. But it is also important to know how long is the process active. This is needed to develop process effectiveness metrics and also to decide for how long should the data be kept in active records. The quest for knowledge about process output, throughput and duration will guide our data mining effort. In practical applications, a business process consists of many elementary processes added together. For instance, the university "production of credit hours" is the sum of the enrollment histories of individual students. Contingency tables can aggregate individual "customer processes" into a business process summary.

For a given cohort of students, the total throughput can be described by the total number of credit hours. A more detailed knowledge on credit hour acquisition can be provided by the histogram of the attribute "credit hours".

Throughput in the next business cycle can be predicted from the regularities discovered in the past data. For instance, there can be a regularity on the total process output in the function of time. It may turn out, however, that the predictions may not be so accurate as those obtained by splitting the process into several subprocesses.

Process can be split into parallel components. Each of the parallel subprocesses uses a part of the input and contributes part of the output or throughput. The inputs and outputs of parallel subprocesses add up to the input and output of the entire process, in analog to the Kirchoff Law for processes of current flow in the electric circuits.

Process can be decomposed sequentially. Process P can be divided into subprocess P_1 followed by P_2, when the output of process P_1 is the input to process P_2. Whenever there is a need for predicting the process input, we can seek explanation of the input to the business process P by processes that are sequentially prior to P and supply parts of P's input. Those data come from various sources, external to the business process database.

Parallel decomposition can be guided by regularities between input and output/throughput attributes. Let C be an attribute which describes the throughput of process P. Let V_C be the range of values of C. The histogram of C is the mapping $h : V_C \longrightarrow N$, where for each $c \in V_C$, $h(c) = n$ is the number of occurrences of c in the data. We can use the histogram of C as a measure of efficiency of the process P, but when the values of C are numbers, we can integrate the histogram to obtain the summary output $T_P = \sum_{c \in C}(c \times h(c))$.

Consider a regularity in the form of statistical contingency table that captures the relationship between an attribute A that describes the input of P and the throughput C. For instance, A can represent the grades received by the students in high school, while C can represent the number of cretit hours taken in college. For each value $a \in V_A$ and each $c \in V_C$, $p(a, c)$ is the probability of the value combination (a, c) derived from data. Using the distribution of inputs given by the histogram $h(a), a \in V_A$ we can compute the throughput histogram, by

converting probabilities $p(a, c)$ into $p(c|a)$ and using the latter in the following way:

$$h(c) = \sum_{a \in A} p(c|a) \times h(a), \forall c \in V_C . \quad (1)$$

When the histograms of A and C and the contingency table have been derived from the same data, the equation (1) is an identity. But when only the histogram of the input A is known for a similar process Q, the probability distributions $p(c|a)$ for each a and equation (1) can provide valuable predictions for the output of Q.

When probability profiles (vectors) $p(c|x), x \in V_A$, differ for different C values, it makes sense to think about process P as a combination of parallel processes P_a for each value $a \in V_A$. We can further decompose each P_a in different ways either in parallel or in sequence. Parallel decomposition is particularly useful when:

1. differences between probability profiles are big;

2. attribute A can be controlled by the business process manager;

3. the histograms of A undergo big variations for inputs made at different time;

4. we expect that process P goes on differently for different P_a.

Let $h_Q(a)$ be the histogram of A for the new cycle Q of the business process represented by $p(a, c)$. Let $N_P = \sum_{a \in A} h(a)$ be the total of input units for P, and $N_Q = \sum_{a \in A} h_Q(a)$ be the total of input units for Q. Using the total throughput T_P of P , the throughput of Q is predicted as $T_Q = T_P \times (N_Q/N_P)$. The decomposition of P into P_a's leads to the prediction of T_Q as $\sum_{c \in C}(c \times \sum_{a \in A} p(c|a) \times h_Q(a))$.

When the conditions 1-3 are satisfied, the difference between both predictions is particularly large, justifying the decomposition into parallel processes.

Sequencial analysis is useful when some attributes describe stages of a process. Often, in addition to the input and output attributes, other attributes describe intermediate results. Suppose that B is an attribute that allows us decompose process P sequentially into P_1 and P_2, and use B as a measure of effectiveness of P_1. This is particularly useful, when the subprocess P_1 applies only to some inputs while some other inputs can be used as a control group, so that the effectiveness of P_1 can be compared with the control group. We will discuss remedial classes and financial aid and their effectiveness as examples of P_1.

Sequential analysis is also important when it leads to knowledge useful in predicting process input. For instance, the number of new students depends on the number of high school graduates, on the cost of study per credit hour, and so on. Regularities derived from past data lead to predictions about new students who will enroll in the future.

Since the attributes that can provide predictions of new input to the business process are typically not available in the business process database **B**, other databases must be searched for relevant information. A relevant database **D** must

include at least one attribute A_1 that provides information temporarily prior to the input A_2 of **B** and at least one attribute J that can be used to join **B** and **D**. Aggregation of data in one or both databases may be required to reach the same granularity of J. Further, the join'ed datatable **B+D** must yield a regularity between A_1 and A_2. In this way a regularity can be found between high school graduation and freshmen enrollment. Since high school graduation figures are known in May, such a regularity provides predictions of college enrollment in August. The information from tables (1) and (3) is joined with data in (5) by the use of attribute "county". Enrollment data must be aggregated to provide the count of students who are residents of each county.

2 An example of process analysis: university enrollment

Understanding the factors in enrollment decline and increase is critical for universities, as often the resources available to the university depend on the number of credit hours the students enroll. 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.

Student databases kept at every university can be instrumental in understanding the enrollment. We have applied the process analysis methodology to a university database exploration and step after step expanded our understanding of the university enrollment process. The initial discovery goals have been simple but their subsequent refinement led to sophisticated knowledge that surprised us and influenced university administrators. Within the limits of this paper we only describe a few steps and a few results that illustrate the business process analysis driven by knowledge discovered in data. Our previous research on enrollment has been reported by Sanjeev & Żytkow (1996).

The data came from several sources. Consider a student database that consists of the following files (tables) 1-4 and (5):

```
(1) Grade Tape for each academic term (Mainframe, Sequential File)
(2) Student History File (Mainframe, VSAM file)
(3) Student Transript file (Mainframe, VSAM file)
(4) Student Financial Assistance (Mainframe, DB2 database)
(5) High School Graduates; Kansas State Board of Education
```

We used temporal precedence to group the attributes into three categories. Category 1 describes students prior to their university enrollment. It includes *demographics:* age at first term, ethnicity, sex, and so forth, as well as *high school information*: the graduation year, high school name, high school grade point average (HSGPA), rank in the graduating class, the results on standardized tests (COMPACT) and so on. All these attributes come from the tables (1) and (5).

The attributes in **Category 2** describe events in the course of studies: hours of remedial education in the first term REMHR, 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, number of times placed on probation, and academic dismissal. All these attributes come from the tables (2), (3), and (4).

Category 3, also referred to as goal attributes, captures the global characteristics of a business process that describe output, throughput and duration. In our example we use academic degrees received (DEGREE) as direct measure of a successful output. Bachelor degrees are awarded after completing approximately 120 credit hours. But all credit hours taken by a student contribute to the total credit hours, which is the determinant of university's budget. Thus the total number of credit hours taken (CURRHRS) measures process throughput. Process duration can be measured in the number of academic terms enrolled (NTERM) by the students. All these attributes come from table (3).

Query-1 prepared the initial data table. In our walk-through example we use attributes in all three categories. We analyze a homogeneous yet large group of students, containing first-time, full-time freshmen with no previous college experience, from the Fall 1986. The choice of the year 1986 provides sufficient time for the students to receive a bachelor's degree by the time we conducted our study, even after a number of stop-outs.

Query-1 prepares data in a number of steps: (a) Select from the table (1) for the Fall 1986 all freshmen **(class=1)** without previous college experience **Prev=0** and **sex=1 or sex=2** for new males and/or new females; (b) join the result with the transcript file (3) by SSN; (c) create the attributes that total credit hours, remedial classes, the number of semesters enrolled, average the grades, etc. (d) project the attributes in Categories 1, 2, and 3, discussed above, including all the attributes created in step (c).

We used the 49er KDD system to search for knowledge. 49er (Żytkow & Zembowicz, 1993) discovers knowledge in the form of statements "Pattern P holds for data in range R". A range of data is the whole dataset or a data subset distinguished by conditions imposed on one or more attributes. Examples of patterns include contingency tables, equations, and logical equivalence. Contingency tables (Bhattacharyya & Johnson, 1986) are very useful as a general tool for expressing statistical knowledge which cannot be summarized into specialized patterns such as equations or logical expressions. Since enrollment data lead to fuzzy knowledge, in this paper we will only consider contingency tables, although some enrollment knowledge has been approximated by equations.

49er can be used on any relational table (data matrix). It systematically searches patterns for different combinations of attributes and data subsets. Initially, 49er looks for contingency tables, but if the data follow a more specific pattern, it can follow-on with a specialized discovery mechanism, such as search in the space of equations.

Statistical tests measure the significance and predictive strength of every hypothesis, which is qualified as a regularity if test results exceed the accep-

tance thresholds for each test. The significance indicates sufficient evidence. It is measured by the (low) probability Q that a given sample is a statistical fluctuation of random distribution. While in typical "manual" applications of statistics researchers accept regularities with $Q < 0.05$, 49er typically uses much lower thresholds, on the order of $Q < 10^{-5}$ because in a single run it can examine many thousands of hypotheses. In a large hypotheses space many random patterns look significant. 49er's principal measurement of predictive strength of contingency tables is based on Cramer's V coefficient

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

for a given $M_{row} \times M_{col}$ contingency table, and a given number N of records. Both Q and V are derived from the χ^2 statistics which measures the distance between tables of actual and expected counts.

Cramer's 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 the distribution expected from 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 to compare regularities found in different subsets and for different combinations of attributes.

Quest-1, the initial discovery tasks, has been a broad search request. We requested all regularities between attributes in Category 1 and 2 as independent variables against attributes in Category 3.

In response to quest-1, the 49er's discovery process resulted in many regularities. In this paper, we focus on a selected few. Let us mention a few other examples (Sanjeev & Żytkow, 1996): big differences exist in college persistence among races and among students of different age; students who changed their majors several times received degrees at the highest percentage.

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 composite ACT score is a better predictor than either high school grade point average (HSGPA) or the ranking in the graduating class. This can be seen by comparing Tables 1-a and -b. Table 1-b shows a regularity which is slightly more significant ($Q : 10^{-34} vs 10^{-32}$) and stronger (V:0.20 vs 0.19).

Analogous patterns of approximately the same strength and significance relate COMP ACT and HSGPA with all three goal variables. The corresponding tables cannot be reproduced due to the space limit.

Table 1. Count tables for HSGPA vs{CURRHRS, NTERM, DEGREE}; COMPACT vs CURRHRS

CURRHRS (a)

120 +	0	11	102	92	73
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
HSGPA	F	D	C	B	A

$\chi^2 = 229, Q = 1.7 \cdot 10^{-32}, V = 0.19$

CURRHRS (b)

120 +	40	78	59	108	5
90-119	40	44	24	38	3
60-89	26	32	28	29	0
30-59	57	68	43	36	0
1-29	262	196	65	56	1
0	38	8	4	2	0
COMPACT	missing	< 19	≤ 22	≤ 29	> 29

$\chi^2 = 221, Q = 6 \cdot 10^{-34}, V = 0.2$

NTERM (c)

12 +	0	10	41	26	8
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
HSGPA	F	D	C	B	A

$\chi^2 = 168, Q = 3 \cdot 10^{-23}, V = 0.2$

DEGREE (d)

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

$\chi^2 = 157, Q = 1.5 \cdot 10^{-28}, V = 0.25$

In conclusion, parallel decomposition of the process by the ACT scores is very useful, and may lead to further findings, when different subprocesses are analysed in detail. We will see that happen in the case of remedial instruction. Since the values of ACT and HSGPA are closely related, it does not make sense to create separate parallel processes for HSGPA.

A sequential analysis problem: does financial aid help retention? Financial aid belongs to Category 2, the events that occur during the study process. It is available in the form of grants, loans and scholarships. The task of **Query-2** has been to utilize the financial aid data. By joining the source table (4) with the table obtained as a result of query-1 we augmented that table with 64 attributes (eight types of financial aid awarded to the students in each of the 8 fiscal years 1987-94) attributes.

Quest-2 confronted the new aid attributes with with our goal attributes (Category 3). The results were surprising. No evidence has been found that financial aid causes students to enroll in more terms, take more credit hours and receive degrees. For instance, the patterns 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 use query-3 and query-4 to select 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 each of two subgroups we tried quest-3 and quest-4 analogous to query-2.

Table 2. Actual Tables for DEGREE vs REMHR (a) all students (b) remedial needing

(a)

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

$\chi^2 = 31.2, Q = 0.0, V = 0.106$

(b)

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

$\chi^2 = 5.06, Q = 0.89, V = 0.091$

In the quest-3 (students needing remedial instruction) we sought the possible impact on Category 3 variables of financial aid received in the first fiscal year, that is, when the remedial instruction has been provided. The results were negative: 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 response to quest-4, 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). These findings indicate that financial aid received by students in the first year was not helpful in their retention.**

Using query-5 we created an attribute which provides the total dollar amount of aid received all fiscal years 1987 – 1994. Now, with quest-5 we sought the impact of this variable on the goal attributes in Category 3. Finally, a positive influence of financial aid has been detected (Sanjeev & Zytkow, 1996), but the results are due to the fact that in order to receive financial aid the student must be enrolled. We could not demonstrate that financial aid plays the role of seed money by increasing the enrollment to years when it hasn't been received.

It is useful to combine data from different datasets. Within one database the join operation can be performed on keys such as social security number, while the joins across different databases may require group values and be preceded by aggregation.

An example of sequential analysis: remedial instruction: One of the independent variables in Category 2, used in query-1 has been REMHR **(total number of remedial hours taken in the first term). An intriguing regularity has been returned by the search (Table 2):** *"Students who took remedial hours in their first term are less likely to receive a degree"*. **The percentage of students receiving a degree decreased from 31% for** REMHR**=0 to 13% for** REMHR**=8. This is a disturbing result, since the purpose of remedial classes is to prepare students for the regular classes.**

Query-6: Select students needing remedial instruction. After brief analysis we realized that Table 2 is misleading. Remedial instruction is intended only for the academically under-prepared students, while students who do not need

remedial instruction are not the right control group to be compared with. In order to obtain relevant data 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 as: (query-6:) 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.

Quest-6: For data selected by query-6 search for regularities between REMHR **and process performance attributes. Use attributes in Category 1 to make subsets of data.** 49er's results were again surprising because no evidence has been found that remedial instruction helps the academically under-prepared students to enroll in more terms, take more credit hours and receive degrees.

Table **??** indicates that taking remedial classes does not improve the chances for a student to persist to a degree. For instance, those students who 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. A similar table indicates no relationship ($Q = 0.98$) between hours of remedial classes taken and number of terms enrolled.

Finally, let us briefly discuss how **external data can be used to predict part of the input.** Table (5) provides the number of high school graduates by county. Knowing the fraction of that number who enroll at WSU, we can predict in June, a part of university enrollment in August. **Query-7** joined tables (1) and (5) by the year, but prior to that it aggregated each table (1) for the corresponding years by the county, counting the numbers of students per county. Now, **quest-7** has determined the percentage of students who transfer from high school to WSU. That number, recently at about 20% can be used to predict new enrollment.

3 Impact of findings on the business process

In 1995-97 the results of our enrollment research have been presented to senior university administrators including *Associate to the President, Vice President of Academic Affairs, Associate Vice Presidents, Director of Budget and academic college Deans*. Many of them chaired or were part of executive purpose committees like *Strategic Plan Task Force, Academic Affairs Management Group, and University Retention Management Committee*.

Starting in the Spring 1997, for the fourth consecutive enrollment period, WSU's enrollment has increased. It is the most consistent enrollment increase in the 1990s. While business decisions are not always based entirely on empirical evidence, such evidence certainly helps to make well-informed decisions. We discuss here some of the strategic decisions, and outline how our findings have formed their underlying empirical foundation.

"WSU will recruit and retain high quality students from a variety of ethnic and socioeconomic backgrounds" is the second out of the five stated *Goals and Objectives* in the draft *Strategic Plan for Wichita State University.* This strategic plan, outlined in 1997, is currently being presented to the various university constituencies, such as the faculty senate, for review and acceptance.

In 1995 and 1996, our research uncovered that academic results in high school are the best predictor of persistence and superior performance in college. Tables 1 show regularities between composite ACT and average grade in high school (HSGPA) as predictors and the college performance attributes: cumulative credit hours taken (CURRHRS), total academic terms enrolled (NTERM), and degrees received (DEGREE).

The eight year graduation rate measure has been included as WSU's **performance indicator.** A strategic planning process called VISION 2020, initiated by the State of Kansas, requires the universities to formulate a set of performance indicators and report the results (Chambers & Sanjeev, 1997). The first of the core indicators concerns *undergraduate student retention and graduation rates.* The report mandated by the Regents asked for graduation rate measures after four, five, and six years. But the students at WSU take usually longer than six years to graduate: they tend to stop-out for several academic terms during their college careers and enroll in less than 15 hrs in one semester. This phenomenon has been a conclusion from our studies in 1995 and 1996. It can be partially observed in Table 1-c. It shows that a significant percentage of students enroll above 11 terms. In addition many students stop out for few semesters. As a result, among the six Regents universities in Kansas, WSU is the only institution in which the graduation rate is also measured at the end of the eight year.

Our negative results increased the awareness of the cost of remedial education. The upcoming replacement of open admission by entry requirements is pushing aside the question of reforming the remedial education, but university administrators are increasingly focusing on the effectiveness of remedial classes. Most recently, in the Fall 1997, a cost study has been conducted on remedial education programs. *Can those costs be justified in the absence of empirically provable success?* is currently discussed by the administrators.

4 Process networks vs. Bayesian networks

While Bayesian networks (Buntine, 1996; Heckerman, 1996; Spirtes Glymour and Scheines, 1993) emphasize the relationships between attributes, the business process networks captures the relations between states of affairs at different time. They resemble the physical approach to causality: the state of a system at time T_1 along with the domain regularities causes the state at time T_2. This is a more foundational understanding of causes.

In distinction to Bayesian networks that relate entire data through probabilistic relations between attributes, a subprocess often involves only a slice of data. For instance, remedial classes are taken by only some students. One part of

a process can be decomposed differently than another part and the corresponding subsets of records can hold different regularities. For instance, a causal relation between attributed may differ or not exist in a subset of data characterized by a subset of attribute values.

5 Conclusions

We have introduced a number of knowledge discovery techniques useful in analyzing a business process. Business processes can be divided into parallel components when it improves the predictive power. Processes can be analyzed sequentially, to find out how the preceding process influences the next in sequence.

We also demonstrated how queries that seek data can be combined with quests that seek knowledge. In a KDD process, queries are instrumental to quests. In addition to explicit queries, the search process "asks" many queries internally.

The data that we used as a walk-through example have been obtained from a large student database, augmented with statistics kept by the State of Kansas. Both have been explored by 49er. In this paper we discuss a few examples of practically important findings that have influenced the University policies.

References

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

Buntine, W. 1996. Graphical Models for Discovering Knowledge, in Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy eds. *Advances in Knowledge Discovery and Data Mining*, AAAI Press, p.59-82.

Chambers, S. & Sanjeev, A., 1997. Reflecting Metropolitan-Based Missions in Performance Indicator Reporting, Metropolitan Universities, Volume 8, No. 3, Winter, p. 135-152.

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.

Heckerman, D. 1996. Bayesian Networks for Knowledge Discovery, in Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy eds. *Advances in Knowledge Discovery and Data Mining*, AAAI Press, p.59-82.

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.

Ohrn, A.; Komorowski, J.; Skowron, A. & Synak, P. 1998. The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets - The ROSETTA System, To appear in *Rough Sets in Knowledge Discovery*, L. Polkowski and A. Skowron (eds.), Physica Verlag, 24 pages.

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

Sanjeev, A., & Zytkow, J., 1996. A Study of Enrollment and Retention in a University Database, Journal of the Mid-America Association of Educational Opportunity Program Personnel, Volume VIII, No. 1, Fall, p. 24-41.

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

Ziarko, W. & Shan, N. 1994. KDD-R: A Comprehensive System for Knowledge Discovery in Databases Using Rough Sets, in T.Y.Lin & A.M. Wildberger eds. Proc. of Intern. Workshop on Rough Sets and Soft Computing, pp. 164-73.

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