IR Applications

Using Advanced Tools, Techniques, and Methodologies

Association for Institutional Research

Enhancing knowledge. Expanding networks.
Professional Development, Informational Resources & Networking

Volume 10, July 19, 2006

Using Academic Behavior Index (AB-Index) to Develop A Learner Typology for Managing Enrollment and Course Offerings – A Data Mining Approach

Jing Luan, Chief Planning, Research and Knowledge Systems Officer, Cabrillo College

Abstract

This exploratory data mining project used distance-based clustering algorithms to study three indicators of student behavioral data collectively called AB-Index, and established a typology of six types of learners for a suburban community college. The study is based on the notion that student behavioral data are a good basis for new ways of doing research studies rather than using non-behavioral data, such as gender or race and intended educational goals. The discoveries from this data mining endeavor are meaningful for understanding and measuring students' behaviors. The study encapsulated and discussed several fresh and novel topics and analytical approaches. The study uncovered previously unknown differences in both output (FTES) and outcomes (GPA, Persistence) across the learner types which may greatly enhance a college's ability to monitor the changes and to make appropriate adjustment to enrollment and teaching strategies. The study noted the lack of predictive power of traditional indicators, such as race or gender, across learner types within the typology. The study also employed several less often used data visualization techniques, such as drop-line charts and the Web graph.

Rationale

Developing typologies is fundamental to science (Bailey, 1994; Fenske, et al., 1999) and is an important research activity (Han and Kamber 2001). Biology would have suffered an irreversible setback had there not been the creation of taxonomies that classified organisms into phyla and species, etc. (Fenske et al., 1999). When confronting large scale databases as researchers increasingly find themselves doing, activities of clustering, classification or grouping are essential. Without first attempting to reduce complicated datasets into manageable pieces, it is difficult if not impossible to completely understand hidden patterns in data. Yet, typological research is underused and under-researched in social science (Luan, 2002) as evidenced by the scarcity of research literature on this subject. The lack of authors who worked on typologies for student behavior created a perceptible gap between what was done and what needs to be done. Astin (1993) conducted an empirical typology of college students in hopes of gaining insights into student life. Fenske et al. (1999) proposed an early intervention program typology. Levine et al. (2001) developed an empirically-based typology of attitudes toward learning community courses. Johnstone in 2004 presented discoveries based on naturally existing student types to understand their various education activities. Zhao, Gonyea and Kuh (2003) theorized student typologies based on students' engagement behaviors in-class and out-of-class.

Typologies can be either qualitative or quantitative. Some qualitatively derived classifications exist in higher education institutions. A university or community college mission statement typically describes the types of courses they offer and the types of students they serve. Categories such as science majors, upper division classes, or career education are often used. These qualitative typologies help describe *who* the students are and *what* they are taught. On the other hand, typologies for the purpose of describing *how* students behave academically are not well established. Behavioral psychology led by B.F. Skinner (Zuckerman, 1994) called attention to the actions of the subjects in addition to the descriptions of the subjects. Not only do behaviors help us better understand the subjects, their malleability provides opportunities to affect change unlike the unchangeable physical background characteristics of students. The practiced method of charting outcomes of success rates or graduation rates by demographic variables

is useful, but outcomes are not necessarily behaviors. Further, continued practice of reporting academic outcomes by gender or race help perpetuate the eponymy by attributing outcomes to physical/cultural traits. In education, we may be too enamored with using what is easy to notice and to record such as subjects' physical traits and demographic characteristics. How do previous behaviors of the subjects relate to their subsequent behavior is the major focus of this research.

Generally speaking, students' behaviors are an inherently more reliable way to gauge if students are achieving their educational goal than their declaration of an education goal on college application forms (Perry, 2005). Holland (1966) reasoned that students, as active agents, made decisions that helped describe who they really were. Pascarella and Terenzini (1991) and Zhao, Gonyea and Kuh (2003) observed that one can learn about students through what they do and that their actions are better predictors of desired college outcomes than their background information. Through classifying students' behaviors rather than their intentions or opinions, clustering algorithms can rely on students' real "actions" to sort them into distinct clusters (groups) based on what students do, not based on what they are or what they say.

These behavioral clusters, once proven to be valid, become a typology and a new class of measures for use in institutional analysis and reporting. Behavioral clusters supplement existing measures such as demographics, major, and even outcomes. The demographics and other qualitative variables are still very useful in cluster analysis. Instead of using them to derive clusters however, they are used for validating and describing the behavioral-based clusters. This study employed this approach of using behavior to form groups of students and then describing them based on their demographic characteristics and opinions.

Typologies are created by identifying clusters. For this research, the clusters will rely on behavioral data. There is a great amount of college data, often in a data warehouse, that are indicative of student behaviors. The first task is to identify data that are most meaningful for establishing typologies. Fortunately, an earlier data mining project at this college resulted in the identification of a group of three sub-indices that collectively are called the AB-Index. The purpose of that project was to supplement existing data points in real-time enrollment reporting at a two-year community college (Luan, 2003).

The AB-Index, defined further in the methodology section of this report, stands for "Academic Behavior Index." This is comprised of three individual indices: The Unit Loading Index, the Adjustment Factor Index, and the Course Volume Index. The Unit Loading Index gives a ratio of units attempted against the total number of courses taken by students. For example, some students average 3 units (credit hours) per course while other students may average 1 unit per course. The Adjustment Factor Index monitors the ratio of number of courses being dropped against the total number of courses enrolled. For example some students may remain in all their courses and have an Adjustment Factor Index equal to zero while other students may withdraw from one or more courses and have an Adjustment Factor Index greater than zero. The Course Volume Index is a count of courses enrolled.

These three behavior-based indices collectively provide good monitoring of the movements and changes among students on a real-time basis because these indices can be obtained directly from a data warehouse that contains current data. Because there are a large number of possible values for all three indices, it becomes a challenge to the human eye when the three indices are examined based on raw data. If "a" is the number of combinations for the Unit Loading Index, and "b" is the number of combinations for the Adjustment Factor Index, and "c" is the number of combinations for the Course Volume Index, there can be almost a*b*c possible combinations! Even though any combination of a, b, and c is one potential type of student behavior, the sheer astronomical number of combinations must be reduced mathematically to reveal key modals of behavior. The best approach to reduce the large number of combinations is to treat the three indices as observations in a spatial environment and use clustering algorithms to identify groupings of observations.

Data mining analyzes data using primarily three approaches: unsupervised, supervised and data visualization. Most people recognize data mining by its names of neural networks, artificial intelligence, and machine learning. These names are typically associated with predictive modeling, also called supervised data mining. Unsupervised data mining, on the other hand, is less popular, but is critically important in understanding data and research subjects. Unsupervised data mining either helps reduce variables or regroup data rows (where variables are in the columns) for the purposes of either uncovering hidden patterns on the variables and/or making the data more manageable. Where supervised data mining examines the variables' abilities to explain the variance for particular "known" output variables, such as graduation, GPA, persistence, unsupervised data mining uses either clustering or *a priori* association analysis techniques to uncover hidden patterns or groupings. Because the purpose of this research is to uncover clusters within a very large number of possible combinations of data points from the sub-indices of the AB-Index, the study relied primarily on unsupervised data mining technique. With research databases increasing in size, it is highly desirable for researchers to conduct unsupervised data mining tasks prior to predictive modeling activities.

Specifically, this study selected software and data mining techniques that had four characteristics. They are: 1) allowing the use of the AB-Index with the 3 sub-indices

directly out of the research data warehouse, 2) allowing the entire data transformation process to stay on one platform tool where the tasks of data cleansing, recoding and deriving could be seamlessly completed, 3) allowing multiple clustering algorithms to be executed all at once to test several dozens of potential cluster scenarios, and 4) allowing the final model to be integrated into the data warehouse for automated and ongoing monitoring of the index.

Research Questions for Data Mining

1) What is an appropriate clustering technique to produce a behavior-based learner typology based on course taking behavior?
2) What is the practical use of the typology?

Research question one is comprehensive in scope. It covers the selection, process, results and validation of a methodology. To answer this question, the study presents the steps in obtaining the clusters and results of intra- and inter-cluster validation as well as face validity1.

Research question two directly determines the value of the typology developed in this study. Data mining can go beyond academic research into immediate application of the knowledge gained; therefore, the second research question addresses the real world application value of the clusters that in the end were considered as a typology of learners.

Design and Methodologies

The research described below includes the following phases: the selection of subjects; data elements; data definitions; data transformations; clustering methods; cluster validation approaches; and cluster face/application validity.

Selection of subjects: The study chose to examine all students (n=15,117) enrolled in Fall 2002 at Cabrillo College - a suburban community college on the west coast with 22,000 enrollment per academic year. It needs to be noted that data recency is not a crucial factor in conducting typological research because the task is to identify abstract types within a particular group of subjects that universally exist in the data.

Selection of data elements: The data elements selection phase resulted in the decision to use the existing AB-Index briefly introduced earlier. The underpinning frame of thought behind the AB-Index was that Students obtain learning from classrooms, but their learning is also a product of many other factors. Financial, social, family obligations and psychological readiness are among a number of influence factors a student uses to manage his/her postsecondary education (Hossler, 1984). Other specific factors may include student employment status, distance to college, stage in life, job requirements, and the college's offerings (Digby, 1986; Rezabek, 1999). Many of these factors that do not necessarily attract the attention of the institution are important to the learner. It is clear that there is no way of getting every possible data point about all of these factors, because college and university data warehouses may be rich in students' academic record history but poor in service and facilities usage data. In addition, few store student perception and attitudinal data that can be linked to individual students in the data warehouse. That is the case in the MIS (Management Information System) system for California's 109 community colleges. However, it is reasonable to expect that the congruent interplay of all these factors would determine a great deal of the way students learn and the ensuing outcome of their learning. Most of these factors eventually manifest themselves as the types of courses students take, the number of courses they take, and the time they take them. These choices are the actions of these students, which become this study's behavioral-based indices.. Therefore, the focus, arguably important, ought to be on the resulting behaviors of the students caused by the influence factors.

The AB-Index includes the number of courses enrolled, the amount of units attempted, and the interesting behavior of course withdrawals. The number of courses in which a student enrolls is the "course volume." It is a summation of all courses taken or dropped by a student in a term. The units attempted is the "unit load." It is a result of calculating the number of credit hours per course taken by a student in a term.

Students will put forward enough efforts until they reach their maximum capacity in managing their course load at which point their option (strategy) is to withdraw from classes. Many studies on class or college retention and dropout identified factors both outside and within the control of the students and the institution (Friedlander, 1980; Rounds, 1984; Windham, 1994). Some studies include biological factors such as gender or race, which presents a controversial and unfair situation for implementing intervention actions. Further, no studies on classifying the dropout/retention oriented behaviors, such as withdrawing from classes, were found. Must withdrawing from classes be considered bad and alarming to college authorities? The fact that a learner withdraws from a class means she/he is reacting to something in their life. Reasonably speaking, the action of withdrawing from a course by a student ought to be viewed as simply and truthfully adjustments a learner makes to his/her studies.

In this study, the action of withdrawing from a course is called the "Adjustment Factor." It is a ratio of courses from which the student withdraws divided by the total courses for which the student enrolled. In order to make it scaled in proportion to the other two indices above for visualization ease, this ratio is multiplied by a factor of 10. The highest possible value for this index is 10 (or 100% withdrawal).

The AB-Index is further defined mathematically below:

- Course Volume (CrsCnt027) = Count of courses taken
- Adjustment Factor (WRationX10) = $\left\{\sum \frac{Ws}{CrsCnt}\right\} \times 10$ (*W is the grade for withdrawing*).
- Unit Loading (UAByCrs027) = Units Attempted / Count of Courses Taken

Note: Units Attempted is the sum of the credit hour value of the courses for which the student initially enrolled for Fall 2002.

As mentioned earlier, the study decided to use only the three sub-indices of the AB-Index for clustering. Other data elements, such as demographics and GPA, were used to study the face and application validity of the clusters. This was an important decision based on a study of developing National Survey of Student Engagement (NSSE) Institution Typologies. That study concluded that using demographic variables to analyze the prospective clusters was superior over using these to generate clusters (Luan, Zhao and Hayek, 2004).

Other fields (variables) identified or calculated for the study include, but are not limited to, the following:

- Full-Time Equivalent (FTES) for both Fall 2002 and the cumulative FTES from previous terms in which the student enrolled.
- Attendance history, the number of terms the student enrolled (TermCnt_Hist).
- Persistence, as defined by students enrolled in Fall 2002 who returned in Spring 2003.
- Grade Point Average (GPA) for both Fall 2002 term and cumulative GPA from all previous terms in which the student enrolled.
- The types of courses, such as transfer, basic skills, vocation education, etc. for Fall 2002.
- Demographic information of gender, race, age (10 ranges), enrollment status, educational goal declared in Fall 2002.

The process of developing clusters can be viewed as being either Agglomerative or Divisive (Bailey, 1994). Agglomerative clustering is sometimes called "bottom up" in the data mining community. Agglomerative clustering starts by allowing each of the N objects to be their own group resulting in N groups. From all these possible N groups, the technique starts combining pairs of groups into more generalized groups. Divisive clustering is considered "top down" (Han and Kamber, 2001). Divisive clustering technique operates from the opposite end by first combining all objects into the smallest number of groups, perhaps just one and then gradually forms new groups by splitting one of the existing groups into two groups. Both agglomerative and divisive clustering techniques are inherently hierarchical because these have to arrive at the final clusters via repeated efforts in an iterative linear sequence of decisions.

Technically speaking, the commonly accepted view of a set of mathematically derived clusters refers to the members or entities in a cluster being maximally similar and members between clusters maximally dissimilar. The more distinct the clusters are, the better the final typology. Because the differences are mathematically driven, clusters are either defined by the centroid-based distance measure $D = \sum \left(\overline{X}_{Ai} - \overline{X}_{Bi}\right)^2$ such as clustering algorithms or by correlational measures (R-analysis), such as factor analysis $f' = \lambda' \sigma_x^{-1}$.

Factor analysis groups data by reducing the number of measures needed to explain the latent relationships between variables and cluster analysis groups data by combining cases. It is rather clear that this study is to group cases, not to reduce variables. Clustering algorithms that are designed to conduct case groupings are therefore chosen.

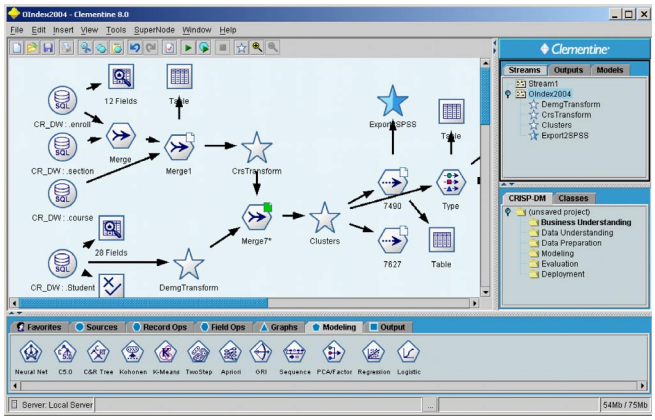
In unsupervised data mining, the researcher does not know how many groups to expect or what the pattern of variables might be. Because no prior known solution exists as a reference point researchers would have to produce many iterations of clusters to find the most appropriate. Each iteration is a cluster scenario.

In actual practice, researchers produce up to 10 clusters for each scenario, large enough a number yet still possible for humans to examine them individually without losing sight of the whole. Based on the Magic 7±2 Doctrine developed by George Miller in 1956, typically 7 clusters, plus or minus 2, are considered appropriate for both understanding the clusters and putting these to practical use. In addition to this rule, most cluster methodologies have analytical metrics that can be helpful in selecting an appropriate number of clusters.

Clementine, a data mining tool by SPSS, was used to carry out every aspect of this unsupervised data mining study. The reason for selecting Clementine is its ability to directly interface with static or live relational databases (as shown in Figure 1), to calculate ad hoc new fields using graphical user interface (GUI) guided nodes, to convert transactional data files into analytical data files, and to allow infinite number of scenarios to be built and examined using its modeling algorithms. All analyses are conducted inside one data stream, which makes it much easier for cross-validation, interpretation, replication and documentation. The following screenshot (Figure 1) illustrates the "data stream" built within Clementine for the entire study, including the nodes used for calculating new fields (variables). In Figure 1, the nodes with the symbol

"SQL" embedded are directly connected to the college's data warehouse[2].

Figure 1: Data Mining Stream for Clustering Algorithmic Bias



Scientific research requires multiple ways to validate its findings. Social science research is no exception. Different algorithms (analytical procedures which are called "modeling nodes" in data mining) were designed for different purposes by different people. As a result, algorithmic bias exists. In data mining model building, algorithmic bias refers to the different interpretable results obtained from a dataset that are caused by using different modeling nodes. For example, even though both the Neural Net node and Kohonen node are based on artificial intelligence theories and approaches, the Neural Net node gives results which are often drastically different from those obtained by using the Kohonen node. The decision-tree-based nodes, such as classification and regression tree (C&RT), approach data mining model building using rule-based information reduction theories and practices. These nodes will naturally work well with some datasets, but not all. Because the results of one algorithm may not fit every dataset, several algorithms should be run on the same dataset to observe the convergence and divergence of results produced by different algorithms. In a recent study, Lei and Koehly (2003) observed the classification errors produced when using linear discriminant analysis (LDA) versus those produced using logistic regression (LR). The authors noted that LDA provided higher accuracy than LR when subjected to the same data. Willett (2004) used both binary LR and C&RT on the same dataset of survey responses to evaluate method bias. After examining the summary statistics produced by the two algorithms, Willett observed that both identified the same variables as most critical, which led him to conclude that the findings were relatively free from method biases. Whether the bias is examined at the summary statistics level or at the individual prediction accuracy level as is often practiced in data mining, the reason is simple. The researcher must be sure that any observed findings are a reflection of the truth in data, not the result of algorithmic differences. This is similar to seeking "a second opinion." An algorithmic bias, if large enough, may result in misguided decision-making.

In this study, to avoid potential algorithmic bias and to choose the best modeling node (the best algorithm), both K-Means and TwoStep clustering nodes were employed to answer the first research question. K-Means is a centroid-based procedure that treats its first case as the starting center point in a Euclidean Hyperspace and continues clustering subjects until a pre-set number of clusters is reached. If a new object in the sequence is too far away, a new cluster is formed. TwoStep uses both log-likelihood and Euclidean distances to determine its cluster centers and it makes two passes through the data. Compared to K-Means, TwoStep allows setting lower and upper limits to cluster numbers as well as the ability to exclude outliers. On the other hand, K-Means produces distance statistics that help explain the relationships of the clusters better than TwoStep does.

Inter-Cluster Validation (Cluster Equity & Cluster Scenarios)

Cluster validation, in this sense, only addresses the mathematical aspect of the clusters but not the content. There are many ways of conducting inter-cluster validation. The first is to examine the membership in each cluster. This is often referred to as cluster size examination for "cluster equity." Cluster size is *a priori* and determining acceptable cluster sizes is subjective (Sun, 2002; Lazarevic, et al., 1999). Further, cluster size and number of clusters also subject themselves to the nature of the data (Han and Kamber, 2001). In some cases oddly small membership in one cluster may indicate the existence of outliers who can be precisely the subjects the data miner needs to find and explore. The decision about how to handle "outliers" will have a major impact on the number and size of clusters because the removal of outliers tends to produce clusters of approximately the same size. This study did not exclude outliers and opted to allow the smallest cluster to be at least 10% of the largest cluster, so that the smallest cluster is not dwarfed into oblivion by the largest cluster. Hence, if one or two clusters is less than 10% of the largest cluster (Smallest Cluster/Largest Cluster)*100), a decision needs to be made whether to keep or discard the entire scenario altogether. This study calls this approach the "cluster equity" rule.

Next is the examination of the number of clusters produced, or "cluster scenarios" validation. The following matrix describes the number of scenarios, or counts of clusters in this study.

Table 1: Clustering Scenario Matrix

	AB-Index
K-Means	4,5,6
TwoStep	4,5,6

Table 1 contains a total of six scenarios, three from using K-Means and three from TwoStep.

How many clusters should there be in order to get closer to the truth? The data mining community tends to follow the notion of Occam's Razor that states the simplest solution may be the best solution. Producing a limited number of clusters helps to ensure parsimony and avoids unnecessarily complex designs. Only when the simple solution would not be satisfactory, would the researcher then move onto something more complex. This principle applies to the selection of variables as well. As additional support for keeping the selection of variables for clustering to a minimum, Han and Kamber (2001) discussed the findings of excessive large number of clusters being generated whenever more variables are introduced. Lazarevic, Xu and Fiez (1999) tested and documented the results of the presence of an abnormally large cluster and onset of several small clusters after introducing three or four new variables that, to a certain degree, caused difficulty in maintaining cluster equity.

Closely related to cluster equity is the analysis of cluster separations. As mentioned earlier, the best clusters are those with the members inside each cluster being maximally similar and members between clusters being maximally dissimilar. All matters of similarities are defined by their mathematical distances in a spatial environment. This also applies to categorical variables that demonstrate asymmetric characteristics. Several methods were developed to conduct this validation, which were collectively called "measures of dissimilarity" (Han and Kamber, 2001). In K-Means it is called the distance to centroid measure. One can also use data visualization to measure dissimilarity by plotting the variables in a 3-D environment. This study utilized the data visualization approach (See Figure 2).

Intra-Cluster Validation (Face and Application Validation)

Face validity is the first step in determining the "content" appropriateness of the clusters through an examination of their distribution of demographic characteristics. The clusters ought to show rather distinctive population patterns to indicate that they have captured meaningful student groups. Another way is to examine the clusters in a 3-D Euclidean space. Because the centers of the clusters are described as the average centroids of the individual cases which belonged to the cluster on the three measures of course volume, units load, and adjustment factor, it is ideal to view the clusters in animation. This study used this method extensively.

The application validity is an extension of face validity and is a very important one. Even after clusters have passed the inter-cluster validation and face validity tests, they may have no practical use. Data mining is most meaningful if it produces actionable information, not just a theory or interesting-to-know findings. Research question two is for the purpose of determining the application value of the clusters, hence uncovering actionable information. The study evaluated cluster application validity by looking at the differences across the clusters on the student outcome variables such as GPA, total units history or term-to-term persistence and on the output variables of FTES and FTES History. These measures are defined in the discussion section. The actual analyses used data visualization and cross-tabulation. The results of application validity produced convincing arguments that the clusters may serve well as a new class of measures for a variety of reporting and decision making needs.

Discoveries

Research question one involves iteratively examining cluster scenarios produced by different algorithms thorough validity analyses. For each cluster scenario, the study conducted both analyses of cluster means and cluster membership as well as visualization of cluster separations. Although both K-Means and TwoStep provided decent cluster membership and separations, clusters built by TwoStep had better delineation when they were analyzed by student GPA. Further, when increased from five clusters to six clusters, TwoStep extracted the extra cluster from Cluster Four without disturbing the rest of the clusters. Because other clusters did not have to be rearranged, this was interpreted to mean that the new cluster represented a meaningful latent dimension. The rest of this study, therefore, relied on the 6-cluster scenario produced by TwoStep[3].

The resulting 6-cluster scenario produced by TwoStep had the following membership as shown in Table 2.

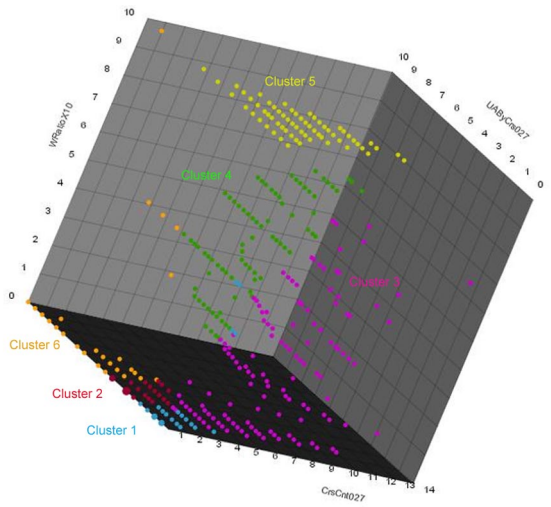
Table 2: Clusters and Membership

	Membership	Small vs. Large Ratio (reference: Cluster Three)
Cluster One	3,025	73%
Cluster Two	3,300	80%
Cluster Three	4,120	--
Cluster Four	2,205	54%
Cluster Five	1,723	42%
Cluster Six	654	16%

The smallest cluster (Cluster Six) was 16% of the largest cluster (Cluster Three). In this case, this cluster scenario seems to meet the inter-validation test – cluster size proportionality.

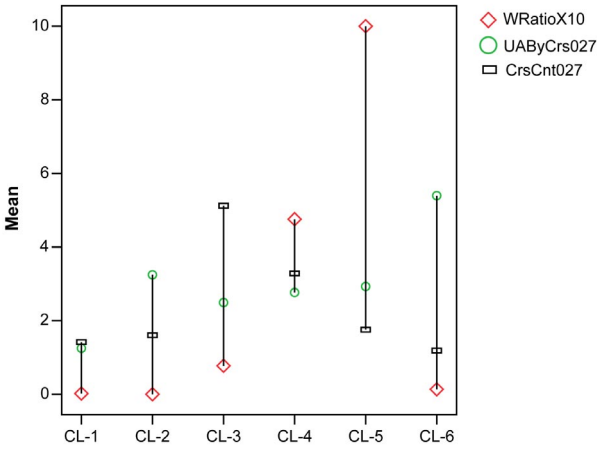
The next task was to produce a 3-D rendition of the spatial separations by using the AB-Index fields and the six clusters. A 3-D graph is most effective in showing the separations of the clusters. The more distinctive each cluster appears in a Euclidean Hyperspace, the better the cluster scenario. The 3-D graph plotted the cases based on their raw scores in three coordinates (vectors) in different colors for identification purposes (Figure 2).

Figure 2: 3-D Graph of the 6-Clusters Generated by TwoStep Using AB-Index (Course Volume, the Adjustment Factor, and the Unit Load)



This 3-D graph, when animated, would show distinctive separations among all clusters. While 3-D charts are best viewed when animated, it is impossible to accomplish it on paper. To accommodate for this deficiency, a drop-line chart substituted for the 3-D chart. Using drop-line chart (Figure 3), a three-dimensional graph was converted to a two dimensional one that is easier to comprehend on paper.

Figure 3: Drop-line Analysis of the Clusters Based on AB-Index



The six clusters in the drop-line chart show distinctive differences coming from the three sub-indices of Adjustment Factor (WRatioX10), Unit Load (UAByCrs027), and Course Volume (CrsCnt027). The following table shows the mean statistics for the three sub-indices.

Table 3: Adjustment Factor, Course Volume and Unit Load by Clusters

		TwoStep Clusters					
		CL-1	CL-2	CL-3	CL-4	CL-5	CL-6
WRatioX10	Mean	.02	.00	.77	4.76	10.00	.14
CrsCnt027	Mean	1	2	5	3	2	1
CrsCnt	Mean	13	15	24	19	14	9
UAbyCrs027	Mean	1.25	3.25	2.50	2.77	2.93	5.40
	Count	3025	3390	4120	2205	1723	654

Note: Course volume in this table is measured by the two variables: CrsCnt027 (students' total number of courses taken in Fall 2002, namely CrsCnt027) and CrsCnt (students' total number of courses taken throughout their academic history at the college).

Students in Cluster One (CL-1) took a small amount of courses, had the smallest unit load and had little or no adjustment. Students in Cluster Two (CL-2), on the other hand, had higher units per course, which is the key difference between students in Clusters One and Two. Students in Cluster Three (CL-3) took more courses (averaging about 5, Table 3), had lower unit per course, and made a few adjustments (averaging 7% of all courses taken, Table 3). Students in Cluster Four (CL-4) appeared to have high adjustment (47% of all courses taken in the fall, Table 3) to the courses they took. Students in Cluster Five (CL-5) dropped all of their courses (100% of their courses taken). They took two courses on average and attempted about three units per course. Students in Cluster Six (CL-6) took the smallest amount of courses of all clusters, but they attempted to get the highest units per course. They managed to drop very few courses (smaller adjustment averaging 1.4% of all courses taken).

In order to assist our analogue human brains with comprehending mathematically derived clusters, clusters were named based on the above observed behaviors of each cluster. By giving names to the clusters, a reader can easily associate the clusters with their demonstrated analogue characteristics. The following table contains brief descriptions of the characteristics (behaviors) and the names given to the clusters.

Demographic variables are valuable to assist with the understanding of the clusters. Using demographics to create clusters would seem to unfairly bring in factors outside the students' control and probably the college as well. On the other hand, using demographics to examine the clusters provides meaningful information. It is almost like giving personalities to the clusters.

Table 4: Cluster Names

Clusters	Characteristics	Name
1	Low on all three sub-indices	Careful Nibblers
2	High on unit load, but low on the other two	Confident Unit Loaders
3	High on course volume, low on unit load and adjustment factor	Well-adjusted Course Packers
4	High on course volume and adjustment factor	Overly Burdened
5	Dropped all courses	Total Withdraw
6	Highest on unit load	Unit Maximizers

Figures 4 through 8 are proportion graphs to display the distribution of background variables, i.e., race and enrollment status, of the students by each cluster. Visually speaking, in Figure 4, the different categories of race appear to be somewhat evenly distributed across the clusters. Figure 5 shows an even distribution (or close to) of the students different enrollment statuses, with Well-Adjusted Course Packers (Cluster Three) and Overly Burdened (Cluster Four) showing slightly more first-time students. Except for the "X" category of gender, which means "Unknown/Unreported," gender is also evenly distributed across the clusters. The graphs of race, gender, and enrollment status, age (Figure 7) and educational goals (Figure 8) show disparity across the clusters.

In Figure 7, a larger portion of younger students are present in the Well-Adjusted Course Packers and Overly Burdened Clusters (Clusters Three and Four). More students chose the goal of Transfer in these two clusters as well (Figure 8).

Figure 4: Clusters by Race

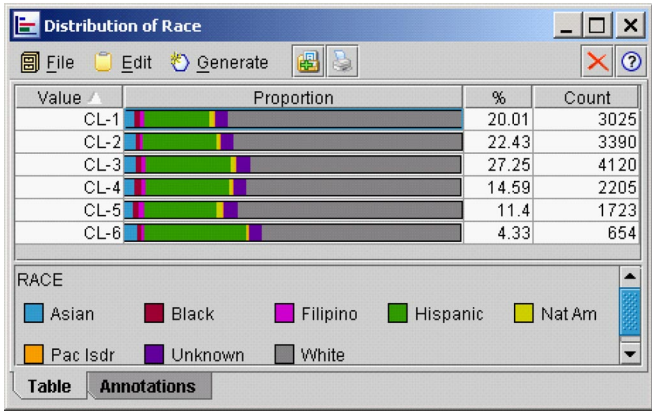


Figure 5: Clusters by Enrollment Status

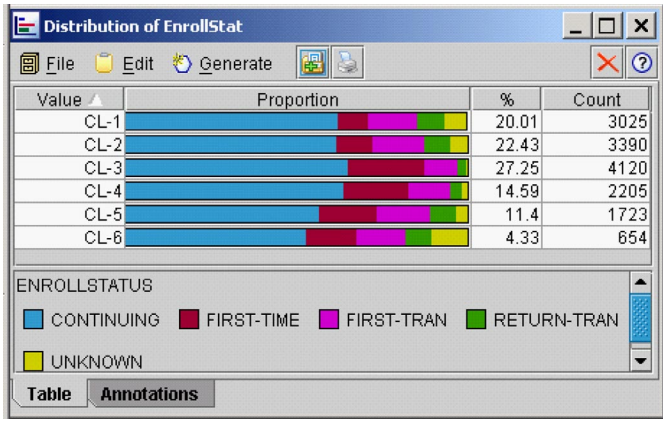
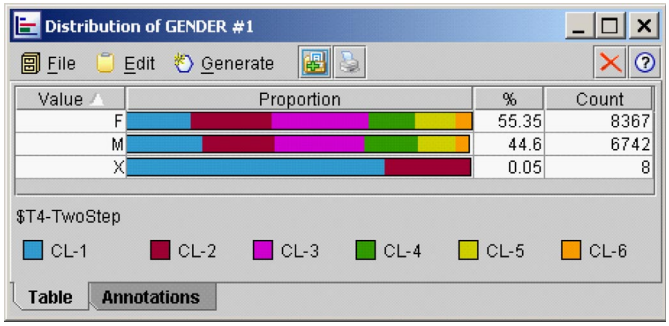


Figure 6: Clusters by Gender (reversed position with clusters)



If using race, gender and/or enrollment status alone, it would appear that the six clusters share similar characteristics. The situation changed drastically when additional demographic variables were introduced.

Figure 7: Clusters by Age

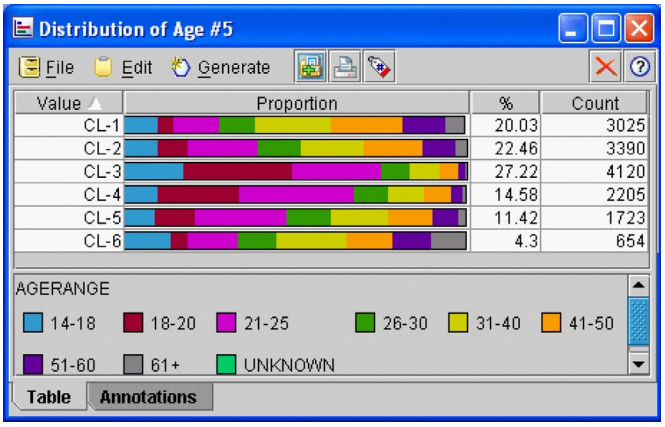


Figure 8: Clusters by Educational Goals

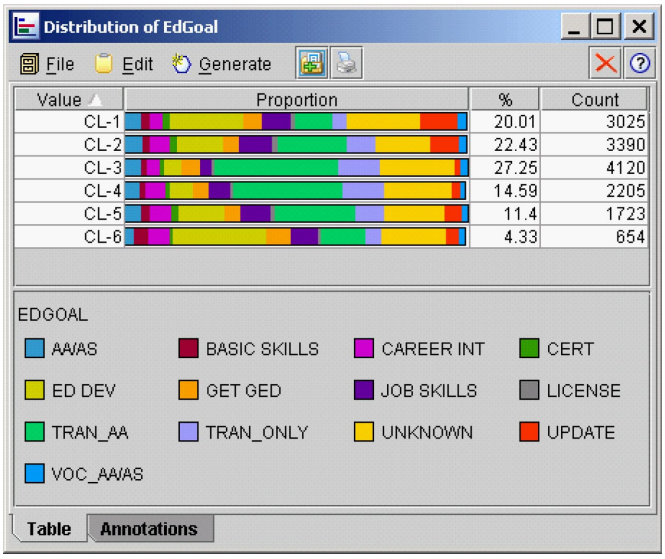
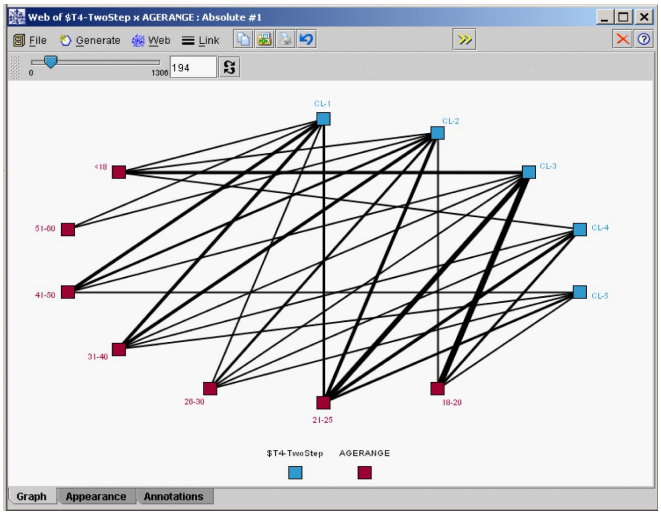


Figure 9: Web Graph of Age and Clusters



Because graphic display of student age and educational goals by cluster showed noticeable differences, the study examined these further. The study first generated a Web graph using age and the six clusters to confirm the differences.

Figure 9 is a special data graph called "Web" graph, which helps indicate the strength of associations between fields (variables). Thicker lines indicate larger proportion of age groups of 18-20, 21-25 in Cluster Three (Well-Adjusted Course Packers)[4].

Chi-square analyses were conducted for age and educational goals separately. Tables 5 and 6 present the parameter statistics for the age groupings (AgeRank) and educational goals by clusters. For age, the differences are significant at .0001 level based on Asymptotic 2-sided (X(45), p< .0001). For educational goals, the differences are also significant (X(60), p< .0001). However, with total n in the thousands, minute differences tend to cause "significance" in a Chi-square analysis (Witte, 1980). The Chi-square analysis conducted here simply validated a visually- based observation. No further analysis using Chi-square is necessary[5].

Table 5: Chi-square Cross Tabulation Parameter Statistic of Age by Clusters

Chi-Square Tests

	Value	df	Asymp. Sig.
Pearson Chi-Square	2994.260[a]	45	.000
Continuity Correction			
Likelihood Ratio	3063.766	45	.000
Linear-by-Linear Association			
N of Valid Cases	15117		

a. 7 cells (11.7%) have expected count less than 5. The minimum expected count is .26.

Table 6: Chi-square Cross Tabulation Parameter Statistic of Educational Goals by Clusters

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1938.442[a]	60	.000
Continuity Correction			
Likelihood Ratio	2011.439	60	.000
Linear-by-Linear Association			
N of Valid Cases	15117		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.35.

The above satisfies the task of addressing research question one that is about the process and results of identifying clusters. The highly iterative process of validating clusters resulted in the selection of a six clusters scenario produced by TwoStep. The clusters have all been appropriately named.

The following portion contains information on the most important aspect of this study that is the real world application value of the clusters. If these clusters provide actionable information about the course-taking patterns of our students then they will prove their usefulness for the college. To prove the usefulness or applicability of the clusters, the study looked at the average values of the output/outcome variables for each cluster. The output variables are FTES and FTES History. An FTES is a full-time equivalent student who has taken 12 units in a term. The FTES History would mean the total number of equivalent FTES a given student produced during his/her entire academic history at the college. This measure is very important in the State of California because colleges are appropriated by FTES. Outcome variables are the typical measures of GPA, Persistence, etc.

The study examined the results by comparing the clusters to each other in terms of the average GPA, persistence (defined earlier) and FTES of their students. Much useful information emerged from this exercise.

The results of comparing the clusters on the output variables of FTES and FTES History are discussed below. While statistics, primarily ANOVA could be computed, this study did not find it necessary, because the purpose of this research is data mining rather than statistical.

Table 7 shows the results of comparing the clusters against the mean (m) of FTES (FTES027) and cumulative FTES (FTES_HIST), as well as the total number of semesters a student enrolled at the college (TermCntBySSN). The Well-Adjusted Course Packers (CL-3) had the highest FTES both for Fall 2002 and for historical cumulative FTES. The reverse is true for the Careful Nibblers (CL-1) and Unit Maximizers (CL-6). However, Careful Nibblers appeared to have the higher number of terms enrolled, which meant this group tended to re-enroll at the college.

Table 7: FTES and Attendance History by Clusters

		TwoStep Clusters					
		CL-1	CL-2	CL-3	CL-4	CL-5	CL-6
FTES027	Mean	.12	.32	.78	.57	.31	.40
FTES_HIST	Mean	1.45	2.54	3.79	3.15	2.14	1.85
TermCntBySSN	Maximum	34	35	31	28	27	18
	Count	3025	3390	4120	2205	1723	654

The results of examining clusters by outcomes variables are discussed below.

Table 8 clearly shows the differences in the mean (m) of the term GPA by each cluster. The Well-Adjusted Course Packers (CL-3) and the Confident Unit Loaders (CL-2) had the highest term GPA. Careful Nibblers (CL-1) and Total Withdraws (CL-5) had the lowest GPA.

Table 8: Term GPA by Clusters

		TwoStep Clusters					
		CL-1	CL-2	CL-3	CL-4	CL-5	CL-6
GPA027	Mean	1.71	2.62	2.67	1.82	.00	2.36
	Count	3025	3390	4120	2205	1723	654

Table 9 shows the results of comparing clusters against the percentage of students in Fall 2002 who returned in Spring 2003 (coded as "YES" in this table). The Careful Nibblers (CL-1), Confident Unit Loaders (CL-2), Overly Burdened (CL-4) and Unit Maximizers (CL-6) had similar persistence rate of a little higher than 60%, while the Well-Adjusted Course Packers (CL-3) had close to 86% of them persisting. On the other hand, Total Withdraws (CL-5), students who withdrew from all courses, had the lowest rate of persisting (29%).

Table 9: Persistence by Clusters

	CL-1		CL-2		CL-3		CL-4		CL-5		CL-6	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
NO	1149	38.0%	1323	39.0%	589	14.3%	751	34.1%	1224	71.0%	240	36.7%
YES	1876	62.0%	2067	61.0%	3531	85.7%	1454	65.9%	499	29.0%	414	63.3%

The results of examining clusters by an output variable of FTES History and other outcomes variables are discussed below.

The following table (Table 10) contains the comparisons of clusters by outcomes variables that were rank-ordered. They were Term GPA, FTES for Fall 2002, and Persistence. The values (the mean of each of the outcome variables) were ranked from 1 to 6 with 6 being the highest. Rank ordered values help with standardizing the performance scales to the extent possible.

What is striking is that the Well-Adjusted Course Packers had the highest performance measured by Term GPA, while the Careful Nibblers had the lowest, excluding the Total Withdraws who had the lowest because of complete withdrawals. Other clusters fell somewhere in between. As measured by FTES, again, the Well-Adjusted Course Packers had the highest FTES generated for the college and they were ranked the highest in Persistence.

Table 10: Clusters by Term GPA, Term FTES, and Persistence

Clusters	Name	Term GPA	FTES	Persistence
1	Careful Nibblers	2	1	Medium
2	Confident Unit Loaders	5	3	Medium
3	Well-Adjusted Course Packers	6	6	High
4	Overly Burdened	3	5	Medium
5	Total Withdraw	1	2	Low
6	Unit Maximizers	4	4	Medium

Legend:
1 through 6 in Term GPA is the rank of GPA with 1 being the lowest.
1 through 6 in FTES is the rank of FTES with 1 being the lowest.
Low, Medium, and High in Persistence denote ordinal categories of Persistence. For example, low persistence would mean lower percentage of students in the cluster that returned to college the following semester.

Table 11 below shows a historical perspective by adding cumulative FTES (FTES History) and Attendance History (total number of terms ever enrolled at the college by the student). It seems that even though the Total Withdraws were low in generating FTES for the term studied (Table 10), they certainly produced high FTES for the college over the years (Table 11). The Overly Burdened also produced high FTES although they appeared to be struggling with their load in Fall 2002. The Careful Nibblers generated very little FTES, but they remained enrolled/ engaged with the college throughout the years.

Table 11: Clusters by FTES History and Attendance History

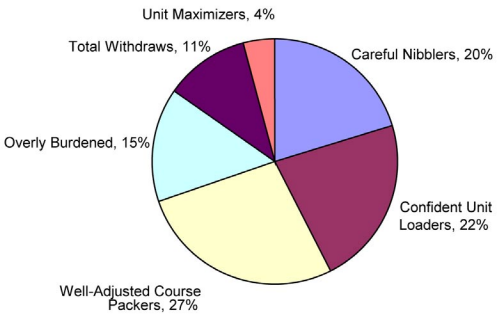
Clusters	Name	FTES History	Attendance History
1	Careful Nibblers	1	5
2	Confident Unit Loaders	3	6
3	Well-Adjusted Course Packers	6	4
4	Overly Burdened	5	3
5	Total Withdraw	4	2
6	Unit Maximizers	2	1

Legend:
1 through 6 in Term GPA is the rank of GPA with 1 being the lowest.
1 through 6 in FTES is the rank of FTES History with 1 being the lowest
1 through 6 in Attendance History is the rank of the total number of courses taken by students in a given cluster throughout their academic history at the college with 1 being the lowest

Summary of Discoveries

The distribution of the six types of learners at the college is shown in the following pie graph (Figure 10). Overall, half of the student population consisted of two types: Well-Adjusted Course Packers (27%) and Confident Unit Loaders (22%). Another 20% were Careful Nibblers. Relatively speaking, this is good news to the college because the types of Total Withdraws, Unit Maximizers, and Overly Burdened with their less predictable enrollment patterns only occupied about 30% of the student body.

Figure 10: Distribution of the Types of Learner



Based on the earlier analysis of clusters by outcomes and by demographics, the author makes the following observations:

- Careful Nibblers were low in every aspect, except for being continually enrolled at the college for many years (having high historical attendance).
- Neither Careful Nibblers nor Confident Unit Loaders produced high FTES, but they have maintained a long history of attending the college.
- Confident Unit Loaders differed from Careful Nibblers by having high term GPA.
- Well-Adjusted Course Packers were by far the stars in every aspect with high term GPA, high FTES, high persistence, and high historical attendance rate.
- Both Well-Adjusted Course Loaders and Overly Burdened had more traditional transfer directed students.
- Overly Burdened appeared to be students who were performing less well as Well-Adjusted Course Packers, even though they are far more alike in demographics compared to other clusters.
- Total Withdraws, the group that totally withdrew their classes from the college, were similar to Confident Unit Loaders in demographics, but demonstrated the tendency to quit completely.
- Unit Maximizers were similar to Careful Nibblers in demographics. They were average in many ways and they had the lowest historical attendance and lowest FTES during their stay at the college.
- Race and gender, the most often used elements for

describing and predicting student behaviors, did not demonstrate their ability to account for much of the variance across clusters.

Discussions

The newly established learner typology has a number of practical implications. These implications are discussed first by the six learner types followed by considering the potential strategies that can be developed for the typology as a whole. In the last portion of this section, significant amount of attention is given to issues and approaches for developing local typologies in institutions.

Careful Nibblers are likely lifelong learning students who live and work within the service community; therefore, they are important constituents of the college for political and bond elections. They have little interest in furthering their studies, yet they do have their distinct needs and prefer to keep their classes at a leisurely manageable level. The same can be said about the second cluster, the Confident Unit Loaders. There are more young students present in this cluster (Figure 7) who may be more likely to become transfer students if their needs are met.

The third cluster, Well-Adjusted Course Packers, are major achievers for a variety of college performance outcomes, such as transfer, persistence, success, even time to degree. The majority are younger students and are focused on transferring (Figure 8).

Those who are Overly Burdened are perhaps students who would have been Well-Adjusted Course Packers, but for some reason, have not done as well academically. It is important to study this cluster closely because they have the potential of transforming into achievers in Cluster Three (Well-Adjusted Course Packers).

Total Withdraws constitute 10% of the total headcount and are troubling because they did poorly in every aspect and severed all relationship with the college – at least for this point in time. Is this a waste of their time and the valuable resources of the college? They did, however, help generate a large portion of the FTES throughout their history at the college. It is a very unique group to say the least.

Unit Maximizers are a group of students who demonstrated a more transitory behavior. They have little enrollment history with the college. When they come, they take a lot of units. It appears that they have high academic potential, but their short stay makes them less easy for the college to establish a rapport with them. How to attract them to stay? Many of them have similar goals of "Educational Development" (Figure 8), as Careful Nibblers do, so perhaps they can move beyond this goal into a higher goal. On the other hand they may be working to advance their career and want to take all of the appropriate courses in a short length of time and then return to their other activities.

Data mining is discovering actionable knowledge and information. What actions can management take in regards to the six types of learners of the behavior-based learner typology? The following decision matrix provides some strategic thinking that includes proposed actions and their associated rationales.

Table 12: Decision Matrix of Proposed Actions and Rationales for the Six Types of Learners

Careful Nibblers	Confident Unit Loaders	Well-adjusted Course Packers	Overly Burdened	Total Withdraws	Unit Maximizers
Percentage: 20%	**Percentage:** 22%	**Percentage:** 27%	**Percentage:** 15%	**Percentage**: 11%	**Percentage**: 4%
Action: Keep offering lifelong learning courses to them.	**Action:** Nudge them to go for more classes or classes with more units.	**Action:** Encourage these "bread and butter" and model students.	**Action:** Provide more targeted counseling.	**Action:** Study why they completely sever their ties with the college.	**Action:** Encourage them to stay.
Rationale: They are happy with periodically taking one or two classes.	**Rationale:** More are younger and are interested in going for more difficult classes (higher units).	**Rationale:** The best performing students who may not need much help, but may appreciate assistance in transferring.	**Rationale:** They seem to be less able to manage course load. They may be a bit too anxious or overly confident.	**Rationale:** A potential huge loss to both the learners and the college. 10% is nothing to be ignored.	**Rationale:** Among all learner types, they took the most units, but had almost the shortest stay.

The proposed strategies for the different types of course taking (Table 12), demonstrates the practical results colleges may achieve by using the learner typology as compared to using students' gender or race alone as sole measures. In addition, research shows that early intervention given to students who have signs of academic stress may help with the overall retention of the institution as well as the success of the individual students (Jefcoat, 1991 and Rudmann 1992). The Overly Burdened students as well as the Total Withdraws may be subjects for a study to observe if some type of intervention in the forms of counseling or tutoring may help reduce their stress level and increase their success. The type of Unit Maximizers with their short tenure at the college should be tracked through national data matching to understand if they tended to demonstrate this behavior everywhere or simply unique to this institution. As a matter of fact, all six types of students ought to be tracked this way to understand their academic life following their study at the current college. This will provide additional insight into the outcomes of the learners by type.

The specific remedies for and approaches to each type of learner may be different from time to time and across different institutions. If the learner typology is developed and implemented elsewhere; however, one essential task is to monitor the percentage distributions of the different

types of learners. Using the pie chart (Figure 10) as a guide, the institutions may set baselines for their own distributions of learners and establish upper and lower boundaries for each type. If the percentage of a certain type rises or falls below a certain limit, it raises red flags for the institution to take actions. For example, if the Overly Burdened increased by 1% each term or the Well-Adjusted Course Packers have reduced by 1 or 2 percentage points, the institution may swing into action to drilldown into that type of learner and to develop an appropriate response.

The significant finding of the less-significant differences among race and gender across the clusters is surprising but not entirely unexpected. Race and a few other demographics are overused in predicting students' outcomes. These will remain important elements to report college data, yet, there is a growing practical difficulty in using these because of increased reluctance of students in stating their demographic characteristics. For example, for the first time in its 2005 enrollment report, the American Council on Education (ACE) added the unknown race category because of the doubling number of students not declaring their ethnicities. Viewed as a whole, demographic types are imperfect proxies for describing, monitoring and managing student learning (Zhao, Gonyea, Kuh, 2003). Therefore, a need for alternative or supplemental indicators is arising as these traditional measures are increasing limitations. One can quickly notice the differences between behavioral-based measures such as the course taking indices and biological/traditional measures, such as those discussed earlier. Behavioral-based research, started by masters of B.F. Skinner and Ivan Pavlov, may very likely play a larger role in social and human subject research in education.

The findings from research question two showed that the statistics of output/outcome data, such as GPA, FTES, or Persistence can be very powerful when analyzed by the typology of student behaviors. Students' biological markers were less informative than their academic behaviors as performance measures. A student of any gender or ethnicity stood an equal chance to be in any of the six types of the learner typology. What set them apart are their behaviors. Their actions directly impacted both output and outcomes of the college.

An open access institution, compared to a selective institution, will have a student body that is naturally diverse in their learning goals, experiences in life, incomes, educational preparedness and even age. The need is to conduct student typology research by first grouping students into meaningful types and then looking at their characteristics. As demonstrated in this study, behavior-based typologies can be very meaningful in helping the college understand what types of students attend. Further, it helps institutions to plan according to the typologies and to learn what to expect based on the typologies.

Several questions tend to be raised in regards to finding optimal clusters, which directly relates to the "goodness" of typologies as well as the application of the knowledge gained from data mining. As many researchers and data miners have stated (Bailey 1994, Berry and Linoff, 2000, Han and Kamber 2001, Sun 2002), there are no perfect clusters because identifying clusters is a subjective activity. It is contingent upon business rules and it is a product of spatial distance measures that are not rule or equation based. In this study, the typology that was developed from the clusters is meant to be used as an alternative to traditional clusters based on measures, such as age, race, major, etc. Both the typology developed by this study and those currently being used do share something in common. In their own right, they are all processes to group subjects for the purpose of describing types of subjects. Thereby embedded in the logic of their commonly shared purpose lies the reason why they need not be perfect.

Some may raise the question about how to institutionalize the newly discovered learner typology. Even though typologies may imply some universal truth—something that maximally separates one group of subjects from the other—all typologies are in essence local and relational. Therefore, the first task is to duplicate the method and approach described in this study to develop one's own institutional typology of learners who take their courses. It is very likely that a similar typology will exist in most two- and four-year colleges. For example, Careful Nibblers are present in most community colleges, but it would seem that there would be fewer of them in a university setting. This is reason an institution should locally develop its own typology.

Efforts to frequently and routinely revise a typology are discouraged. Once the learner types within a typology are identified and properly named, it is not necessary to frequently redo the analysis, just like we do not need to redefine majors or reword categories of educational goals. Instead, institutions should classify incoming students into the typology to monitor change. This is often referred to as segmentation, which uses the patterns discovered from the clusters to classify future subjects. This leads to one additional step not discussed in this study—that is to use a rule-based algorithm to classify new subjects into the typology. This can easily be done using supervised modeling nodes such as C5.0, C&RT, and even Neural Networks. Without a data mining tool, institutions can also use existing regression algorithms within SPSS or SAS to classify (segment) incoming students. Unfortunately using a non-data mining application may cause one to lose some of the ability to easily interface with a live data warehouse, which provides day-to-day monitoring of student activities.

In order to observe the subjects outcomes within each typology, the study used end-of-term data. In developing an actual typology, researchers may want to use point-in-time data so that they can actively monitor students adding and dropping classes/units, thereby coming in and out of different types of learners within the typology.

In addition to the step for a supervised modeling node to classify future students into the six types of learners, other data mining activities, which are feasible but which are beyond the scope of this study, include the following:

1) Drilldown to the dynamics inside the clusters to study the distribution of students by other indicators of the institution.
2) Evaluate the use of other behavior related variables for developing clusters.
3) Conduct predictive modeling guided by the clusters to more accurately determine the accuracy of predicting students GPA, Persistence, and FTES.

Conclusion

Databases are dramatically increasing in size and this size is a challenge for researchers who want to understand hidden patterns in data. This challenge can be met by resorting to some type of high powered data analysis approach. Although data mining is not yet widely adopted because of cost and requirement of both IT and statistics skill, the availability of data mining-based research as identified by Serban and Luan (2002) is gradually improving. This research should help add some practical evidence to support the use of data mining.

This exploratory data mining project used distance-based clustering to cluster the sub-indices of an academic behavior index (AB-Index) and selected a 6-cluster scenario following an exhaustive explorative study of 4-, 5-, and 6-cluster scenarios produced by K-Means and TwoStep algorithms.

The knowledge discovered in this study through the lenses of six learner types offers some critically useful enrollment management values. For example, the existence of a significant portion of learners who have demonstrated a low course taking intensity turned out to be people who are intensely interested in maintaining a connection with the college through periodic enrollment in classes. Attempts to either nudge them to take more courses or to neglect them might result in potential loss of community support. The somewhat shocking presence of 10% of total withdrawals also ought to jump prominently onto management's radar screen. The withdrawals of these students may reflect more than just lost opportunities for the learners. It might also reflect a decrease in overall college accountability and a reduction of resources.

A significant portion of students overburden themselves by packing a large number of courses and it seems that these students ought to be counseled early to better manage their course load so that they could more likely achieve success without too many interrupted attempts. The majority of the student body (70%) consisted of the most welcome (in the eyes of the college) learner types: Well-Adjusted Course Packers and Confident Unit Loaders as well as Careful Nibblers. This should give the college comfort in knowing that they only need to concentrate on the remaining 30%.

This study encapsulates and discusses several fresh and novel topics. In regards to technique, the entire study used data mining for its quantitative analysis. It provided a methodology that produced results. It used a methodology that directly interacted with a data warehouse and provided for multiple algorithms that could be applied on the same data stream. This allowed for considering algorithmic bias, conducting validity examination and generating final clusters.

The study relied on clustering algorithms, a type of unsupervised data mining for which the outcomes are not known *a priori*. In regards to content, the study used academic behaviors of students, called Academic Behavior Index (AB-Index) to produce a learner course taking typology of six types that may lend themselves to be additional measures for reporting and decision making. In regards to the value of applying the resulting typology, the study uncovered previously unknown differences in both output (FTES) and outcomes (GPA, Persistence) across the learner types. Understanding these differences can help enhance a college's ability to monitor the changes in student behavior and make appropriate adjustment to the college's enrollment and teaching strategies. In regards to analysis, the study employed several rarely used data visualization techniques, such as drop-line charts and Web graph. As an additional contribution to the search for better proxies for measuring learning, the study noted the lack of predict power of traditional indicators, such as race or gender, across learner types within the typology.

Editor's Comments

The advent of the computer has been one of the most pervasive and ubiquitous shifts in the short existence of institutional research. From the first discovery of BMD and IMSL we started being able to do analyses that were previously virtually impossible. This has continued as a major trend and influence in both the business we do and the way we do business.

This IR Application by Jing Luan exemplifies both what we can now do and how we can now do it. The major tool he uses is Clementine by SPSS. The data handling is an integrated stream that includes the basic steps of selection of subjects and construction of variables. The analysis focuses on cluster analysis. The interpretation of results uses a combination of statistics and graphs.

The use of exploratory descriptive analysis, of which cluster analysis is only one of the options, has the ability to look into an empirical environment. Because of this empiricism cluster analysis has been attacked because of the potential of misuse in the absence of a conceptual framework. One question that needs to be raised is the adequacy of the conceptual framework surrounding the three indices that Luan uses as the basis of his analysis. If you accept his argument that these are the essence of

enrollment in an environment of open-enrollment, then you would be equally pleased in the environment where student enrollment is the major source of revenue.

A second issue this article raises is the question of complexity. It is evident that the methodologies in data mining are complex. Luan's use of different graphics and his naming of the groups are both good ways to simplify the presenting of the results. The question that is raised is whether there is a simpler way of digging into the behavior of these students. Unfortunately there does not seem to be an obvious alternative.

One overriding premise by Luan, about which there is likely very little disagreement, is that it is better to categorize individuals by what they do rather than by their characteristics. There are several additional methodologies in Clementine, branching algorithms and neural nets. We look forward to future installments by Luan demonstrating the use of these tools.

Special Acknowledgment

In the course of over a year, the study received enormous assistance both in the articulation of findings (discoveries) and the elucidation of discussions from **Dr. Chun-mei Zhao**, Research Scholar, the Carnegie Foundation for the Advancement of Teaching. Her tireless efforts and superb scholarly skills have made the article ever more lucid and worthy. A thank you also goes to my staff members: **Dr. Richard Borden**, Research Analyst, for his winning arguments in designing the AB-Index and **Judy Cassada**, Research Specialist, for proofreading numerous drafts

.

Bibliography

ACE (2005) Annual ACE Report Shows Minority College Enrollment Continues to Climb, but Gaps Still Persist. ACE Press Release, February 14, 2005.

Astin, A. (1993). An empirical typology of college students. *Journal of College Student Development*, 34, 36-46.

Bailey, K. (1994). Typologies and taxonomies: An introduction to classification techniques. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-102). Thousand Oaks, CA: Sage.

Berry, M & Linoff, G. (2000). Mastering data mining: The art and science of customer relationship management (2nd ed.). John Wiley & Sons.

Digby K. E. (1986) A Study of the Factors Which Influence Adult Enrollment in a Technical Institute. ERIC (ED273319).

Fenske, R., Keller, J., & Irwin, G. (1999). Toward a typology of early intervention programs. *Advances in Education Research*. Vol. 4.

Friedlander, J (1980). An ERIC Review: The Problem of Withdrawal from College Credit Courses. *Community College Review*. ERIC (EJ237110).

Han. J. W., Kamber, M. (2001). Data Mining: Concepts and Techniques. Academic Press. San Francisco.

Hossler, D. (1984.) Enrollment Management – an integrated approach. The College Board. New York, New York.

Johnstone, R. (2004). Foothill College Behavioral Segmentation – RP Group Conference Proceedings, Palm Beach, California.

Jefcoat, H. G. (1991). Advisement intervention: A key strategy for a new Age – Student Consumerism. ERIC (ED340546).

Lazarevic A., Xu X. W., Fiez, T., et al (1999) Clustering-Regression-Ordering Steps for Knowledge Discovery in Spatial Databases. International Joint Conference on Neural Networks (IJCNN'99), July 10-16, 1999, Washington, DC.

Lei P. W. & Koehly, L. M. (2003) Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case"). *The Journal of Experimental Education* (Vol.72, No.1, Fall 2003, pp. 25-49).

Levine, J. Jones, P. & Williams, R. (2001). Developing an Empirically Based Typology of Attitudes Toward Learning Community Courses. Part of Presentation for the AAHE Assessment Conference. Denver, Colorado.

Luan, Jing (2003). Developing Learner Concentric Learning Outcome Typologies Using Clustering and Decision Trees of Data Mining (The OIndex report). Presentation at 43rd AIR Forum, Tampa, Florida. May 2003.

Luan, J. (2002). Mastering Data Mining: Predicative Modeling and Clustering Essentials. AIR Forum Workshop Manual. AIR 2002. Toronto, Canada.

Luan, J., Zhao, C. M. & Hayek, J. (2004). Exploring a New Frontier in Higher Education Research: Using Data Mining Techniques to Create an Institutional Typology. Presentation at CAIR 2004 Conference. Anaheim, CA.

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 1956, vol. 63, pp. 81-97.

Perry P. (2005). AB1417 District Performance Framework Project. Presentation at Ohlone College. Fremont, California.

Rezabek R. J. (1999). A Study of the Motives, Barriers, and Enablers Affecting Participation in Adult Distance Education Classes in an Iowa Community College. ERIC (ED463774).

Rounds, J. C., (1984). Attrition and Retention of Community College Students: Problem and Promising Practices. ERIC (ED242377).

Rudmann, J. (1992). An Evaluation of Several Early Alert Strategies for Helping First Semester Freshmen at the Community College and A Description of the Newly Developed Early Alert Retention System (EARS) Software. ERIC (ED349055).

Serban, A. M., Luan, J. (Eds.). (2002). Knowledge management: Building a competitive advantage in higher education: *New Directions for Institutional Research* #113. San Francisco, CA: Jossey Bass.

Sun, Y. (2002). Determining the size of spatial clusters in focused tests: Comparing two methods by means of simulation in a GIS. *Journal of Geographical Systems* (2002) 4:359-370.

Windham, P. (1994). The Relative Importance of Selected Factors to Attrition at Public Community Colleges. ERIC (ED373833).

Willett, T. (2004). Assessing the Digital Divide: an extension of the pulse of the community survey. Galivan College Research Publication.

Witte, R. (1980). Statistics. Holt, Rinehart and Winston, New York.

Zhao, C. M., Gonyea, R. M., Kuh, G. D. (2003). The Psychographic Typology: Toward Higher Resolution Research on College Students. Paper presented at 43rd AIR Forum, Tampa, Florida, May 18-21, 2003.

Zuckerman, M (1994) Behavioral Expressions And Biosocial Bases of Personality. Cambridge University Press. New York, NY.

Endnotes

[1] Face validity is a visual examination of a few key indicators of interest to get an impression if the test appears valid. It is not a sole measure and often precedes more thorough and scientific examination. For example, in a cluster analysis, a visual examination shows one cluster appears to be disproportionaly large and also contains cases that are outliers, it is clear that clusters do not pass face validity.

[2] K-means and TwoStep are also available in SPSS base. Among the key differences between SPSS and Clementine is Clementine's ability of using data stream to develop models and to deploy the model directly over a data warehouse.

[3] Readers herein would not be subjected to the horrific task of sifting through massive amount of tables and charts in order to identify the final set of clusters.

[4] Through sliding the thickness bar (an option in Clementine), more detailed information is revealed that helped further understand what is conveyed in Figure 7.

[5] Additional tables and statistics can be obtained by contacting the author.

IR Applications is an AIR refereed publication that publishes articles focused on the application of advanced and specialized methodologies. The articles address applying qualitative and quantitative techniques to the processes used to support higher education management.

Editor:
Gerald W. McLaughlin
Director of Planning and Institutional Research
DePaul University
1 East Jackson, Suite 1501
Chicago, IL 60604-2216
Phone: 312/362-8403
Fax: 312/362-5918
gmclaugh@depaul.edu

Managing Editor:
Dr. Terrence R. Russell
Executive Director
Association for Institutional Research
222 Stone Building
Florida State University
Tallahassee, FL 32306-4462
Phone: 850/644-4470
Fax: 850/644-8824
air@mailer.fsu.edu

Authors can submit contributions from various sources such as a Forum presentation or an individual article. The articles should be 10-15 double-spaced pages, and include an abstract and references. Reviewers will rate the quality of an article as well as indicate the appropriateness for the alternatives. For articles accepted for *IR Applications*, the author and reviewers may be asked for comments and considerations on the application of the methodologies the articles discuss.

Articles accepted for *IR Applications* will be published on the AIR Web site and will be available for download by AIR members as a PDF document. Because of the characteristics of Web-publishing, articles will be published upon availability providing members timely access to the material.

Please send manuscripts and/or inquiries regarding *IR Applications* to Dr. Gerald McLaughlin.