Learner Typologies Development Using OIndex and Data Mining Based Clustering Techniques

Jing Luan, Ph.D.
Chief Planning, Research & Knowledge Systems Officer
Cabrillo College

Contacting the author
Jing Luan, Ph.D.
Chief Planning, Research & Knowledge Systems Officer
Cabrillo College
6500 Soquel Drive
Aptos, CA 95003
831.477.5656
jiluan@cabrillo.edu

Learner Typologies Development Using OIndex and Data Mining Based Clustering Techniques

ABSTRACT

This explorative data mining project used distance based clustering algorithm to study 3 indicators, called OIndex, of student behavioral data and stabilized at a 6-cluster scenario following an exhaustive explorative study of 4, 5, and 6 cluster scenarios produced by K-Means and TwoStep algorithms.

Using principles in data mining, the study followed a proven data mining process that proceeded from identifying the research questions, to staging the data, to data auditing, and to building scenarios. All scenarios were subjected to data visualization, and in cases appropriate, Chi-square analysis.

This study established 6 typologies of students enrolled at a suburban community college. The study is based on the notion that student behavioral data are good candidates for new facets of research studies, compared to using non-behavioral data, such as gender or race. The discoveries from this study emerged as both meaningful for understanding and measuring students' learning as well as actionable for decision making. The typologies may be added to existing educational strategies for both management and assessment of learning.

Learner Typologies Development Using OIndex and Data Mining Based Clustering Techniques

RATIONALE

Typology is fundamental to science (Bailey, 1994; Fenske, et al., 1999) and is seriously underused and under-researched in social science (Luan, 2002). 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. All in all, only a handful of authors have worked on this subject, which has created a perceptible gap between what has been done and what needs to happen.

Technically, the commonly accepted view of a set of typologies for a particular subject area refers to its members or entities in a group that are maximally similar, and members between groups maximally dissimilar. The more distinct the groupings the better the typologies. Differences are mathematically driven. They 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 approaches grouping data by collapsing fields (variables) and cluster analysis by collapsing cases.

Typologies already 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. These are qualitative typologies that help describe *who* their students are.

On the other hand, typologies for the purpose of describing *how* their students do academically are conspicuously missing. It is, therefore, the focus of this research. As a rule of thumb, students' actions are inherently more reliable as a gauge of their educational goal than the education goals they declare on college application form. Further, what determines learning outcomes is intimately related to what students do, not who they are. Through classifying the behaviors of the students, clustering algorithms rely on the real "actions" of the subjects to sort them into distinct groups for the purpose of proving the notion that what defines a student is what s/he does, not who s/he is or what s/he says.

To develop typologies is to identify clusters. To identify clusters is to use existing or to-be-computed data points or indicators. Which indicators would be ideal candidates as indices for clustering? How would one determine what constitutes a good cluster or clusters? How would one apply the clusters as typologies for higher education? These questions have been converted into the following research questions for the study to address.

In an open access institution, compared to a selective institution, its student body is naturally diverse in their learning goals, experiences in life, incomes, educational preparedness and even age. Analysis and educational treatment cannot be universally applied without encountering high likelihood of failures. The need is present to first group students into meaningful types.

RESEARCH QUESTIONS (DATA MINING GOALS)

1) What are potential indices for clustering analysis?
2) What hidden patterns can be discovered by clustering these indices?
3) What is the practical use for applying these typologies?

DESIGN

There is a great amount of data points in a college data warehouse that are indicative of student behaviors. Therefore there can be just as many indictors of student behaviors that are being tracked by colleges. The task is to identify those that are most meaningful for this research.

The underpinning frame of thought for identifying the indexes for this study is as follows. 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 determine his/her postsecondary education (Hossler, 1984). It is clear that there is no way of getting every possible data point about all the potential factors. However, it is reasonable to expect that the congruent interplay of all factors would determine a great deal of the way students learn and the outcome of their learning. The focus, then, is on the resulting behaviors of the students caused by the influence factors.

To a student, learning is a purposeful process broken down into semesters. S/he acquires learning on a semester by semester basis. Many factors that do not necessarily attract the attention of the institution are important to the learner. For example, their family obligations, employment status, financial capacity, distance to college, stage in life, job requirements, and offerings of the colleges all come into play. Most of these factors manifest themselves as the types of courses they take, the number of courses they take, the time they take them. These choices are the actions they take, which become natural behavioral based indices for this study.

The study chose the number of courses they take, the amount of units they attempt, and an interesting behavior of course withdrawals. The number of courses taken by a student is the "course volume" and is named as "CrsCnt027" in this study to denote the course volume of fall 2002. It is a summation of all courses taken by a student. The units attempted is the "unit loading" and is named as "UAByCrs027" in this study. It is a result of calculating the number of units per course taken by a student.

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 class. Even though the detailed reason why a particular learner withdraws from a class remains a long lasting research and academic debate, the fact that a learner withdraws from a class means s/he is reacting to something in their life. The action of withdrawing from a course is then viewed as the adjustments a learner makes to his/her studies. The action of withdrawing from a course is called the "adjustment factor" and is named as "WRatioX10". It is a ratio of withdrawals and the total course volume. In order to make it scaled in proportion to the other two indices above, this ratio is multiplied by a factor of 10. The highest possible value for this index is 10 (or 100% withdrawal).

These three indicators are deemed as potential indices for clustering. They are collectively called the OIndex in this study. There is no particular reason why it is named OIndex. They are further defined mathematically below:

- Course Volume (CrsCnt027) = Count of courses taken
- Adjustment Factor (WRationX10) = $\left\{\sum \frac{Ws}{CrsCnt}\right\}$ x 10
- Unit Loading (UAByCrs027) = Units Attempted / Count of Courses Taken

The study chose to examine all students (n=15,117) enrolled in fall 2002 at a suburban community college on the west coast with 22,000 enrollment per academic year.

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.

- The number of terms the student enrolled (TermCnt_Hist).

- Persistence, as defined by those enrolled in fall 2002 who returned in spring 2003 (Return).

- 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 (CrsType) for fall 2002.

- Demographic information of gender, race, age (10 groups), enrollment status, educational goal declared in fall 2002.

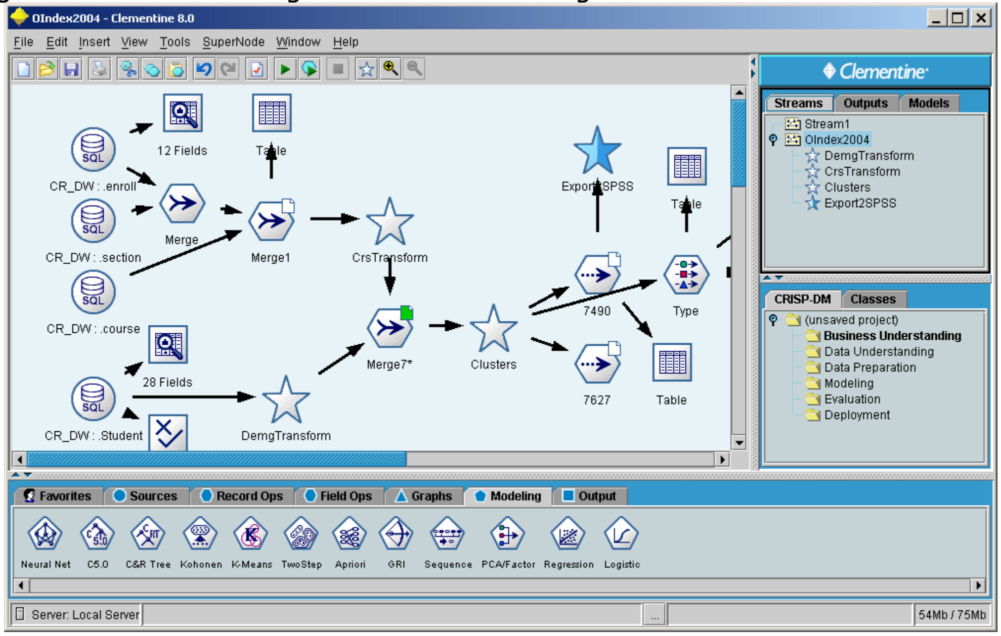
These non-OIndex fields were used for understanding the generated clusters. For example, student demographics and outcome fields were cross-tabulated against the clusters to discover cluster characteristics and dynamics.

METHOD – (KNOWLEDGE DISCOVERY PROCESS)

There are three major types of data mining: unsupervised, supervised and data visualization. Most people recognize data mining by its spectacular names of neural networks, artificial intelligence, and machine learning that are typically associated with predictive modeling, also called supervised data mining. Unsupervised data mining is less known, but is critically important in understanding the data and subjects. Instead of examining the abilities of fields that explain the variance for particular "known" output fields, such as graduation, GPA, persistence, unsupervised data mining use either clustering or apriori association analysis techniques to undercover hidden patterns or groupings. This study employed unsupervised data mining technique.

This study chose to employ data mining techniques largely due to the three reasons: 1) the OIndex with the 3 sub-indices reside in a large data warehouse, 2) there needs to have a tool that allows the tremendous amount of data transformation done in a consistent flow, preferably object based, and 3) the ability to use multiple clustering algorithms all at once to test several dozens of potential clusters. Although data mining is not widely adopted due to cost and requirement of both IT and statistics skill sets, the scarcity of data mining based research as identified by Serban and Luan (2002) is gradually changing. Clementine, a data mining tool bench is used to carry out every aspect of this unsupervised data mining based study. The following screen shot illustrates the data stream built within Clementine for the entire clustering study, including the nodes used for calculating new fields (variables). The reason for selecting Clementine is based on its ability to directly interface with static or live relational databases, to calculate new fields using GUI guided nodes, to convert transactional data files into analytical data files, and to allow infinite number of scenarios being built and examined using its 12 modeling algorithms. All analyses are conducted inside one stream, which makes it much easier for cross-validation, interpretation, replication and documentation.

Figure One. Data mining stream for clustering.



Generating clusters (typologies) requires thorough understanding (domain knowledge) of the input fields (variables). Domain knowledge about the background of the fields and inter-field correlation is important. Lack of such knowledge would result in either meaningless or unnecessarily complicated clusters. It is also advisable to use at least two algorithms and to produce up to 10 scenarios each. A scenario is a collection of clusters produced one time by a clustering algorithm. An exhaustive approach like this helps compare and contrast the differences within clusters and across cluster scenarios. Typically 7 clusters, plus or minus 2, are considered appropriate for both understanding and practical use.

There are many ways of validating clusters. The first is to examine the membership in each cluster. 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. 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.

Other ways are to examine clusters in a 3-D Euclidean Hyperspace, since the clusters are the products of centroid of the individual cases positioned by three coordinates: course volume, units loading, and adjustment factor. This study used this method extensively. Linking the clusters to other fields, particularly outcomes or demographics would also help determine the validity (both face and research validities) of the clusters. Linking can use either data visualization, or cross-tabulation, or some other valid measures.

FINDINGS (DISCOVERIES)

As explained earlier in the design section, action based data that are related to academic activities are well documented in databases. These data points are considerably more stable and accurate, compared to elements such as race, gender or educational goals. Case in point, as high as 20% of the students do not state their race. These and other reasons answered the first research question, that the study should select the three behavioral indices for clustering analysis.

To answer the second research question about hidden patterns to be discovered by clustering, the study used clustering algorithms of K-Means and TwoStep. K-Means is a centroid based that treats its first case as the first centroid in a Euclidean Hyperspace and continues clustering subjects until a pre-set number of clusters is reached. 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 both pre-set cluster numbers or automatic determination of clusters. On the other hand, K-Means produces distances statistics that help explain the clusters better than TwoStep can. The study also took advantage of the scenario building feature of Clementine, which allows infinite numbers of scenarios for each of the modeling algorithms to be built on the same workbench, so that the researcher can compare and contrast the scenarios and pick the best one. The following matrix describes the number of scenarios built by Clementine:

Table 1: Clustering Scenario Matrix

	OIndex	OIndex+Course Types
K-Means	4,5,6	4,5,6
TwoStep	4,5,6	4,5,6

The last column in Table 1 shows clusters built with the 3 sub-indices of the OIndex (course volume, adjustment factor, and unit load) plus the addition of the types of courses taken by the students. The idea was to examine how it would help refine the clusters by introducing the counts of transfer courses, basic skills courses, vocational courses into the equation. However, the clusters produced by adding these elements were too cumbersome to understand and therefore it was abandoned. This is typical in data mining.

The study utilized three methods to validate the clusters produced by the two different algorithms: 3-D visualization of cluster separations, cluster membership, face validity of generated clusters by student demographics.

The study conducted both analysis of cluster means and membership as well as visual analysis of cluster separations for each scenario. Although both K-Means and TwoStep provided descent cluster membership and separations, clusters built by TwoStep had better delineation when they were linked to GPA. Further, when increased from 5 clusters to 6 clusters, TwoSteps extracted the extra cluster from cluster 4, which means it did not disturb the rest of the clusters. The rest of this study relied on results from using only the OIndex and the 6-scenario produced by TwoStep.

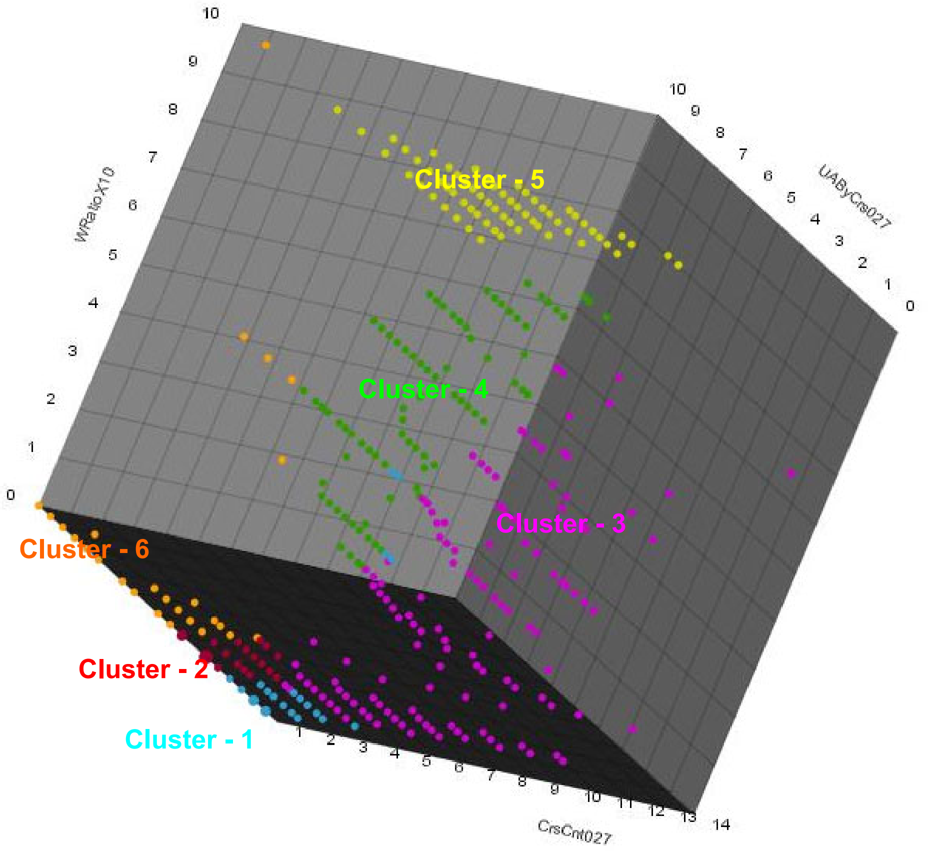
The resulting 6-scenario clusters produced by TwoStep had the following membership:

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%

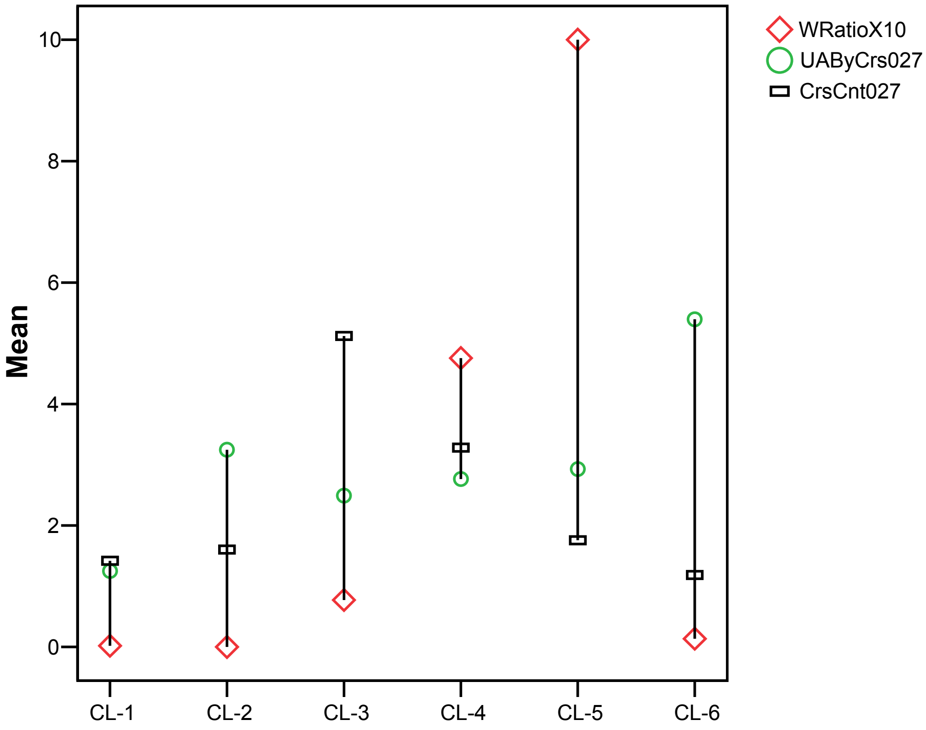
The first task was to produce a 3-D rendition of the special separations by using the OIndex fields and the resulting clusters. The 3-D rendition, technically, is to plot the cases based on their raw scores in 3 coordinates (vectors) and colored by the 6 clusters (Figure 2).

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



Using drop-line chart (Figure 3), a three dimension graph can be converted to a two dimension one that is far easier to comprehend, particularly on paper when 3-D rotation is not possible to demonstrate. As a matter of fact, readers are encouraged to read the 3-D graph first in order to get a sense of where each of the clusters fall in the Euclidean Hyperspace. Please note this drop-line chart used the mean (μ) of each OIndex field.

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



The six clusters show distinctive differences in the three sub-indices of Adjustment Factor (WRatioX10), Units Load (UAByCrs027), and Course Volume (CrsCnt027). Students in Cluster One (CL-1) took a small amount of courses, had the smallest unit load and had little or no adjustment. The only difference between students in Clusters One and Two is that Cluster Two students attempted more units per course. Students in Cluster Three took more courses (averaging about 5), had lower unit per course ratio, and made a few adjustments (averaging 7% of all courses taken). Students in Cluster Four appeared to have high adjustment (47% of all courses taken) to the courses they took. Students in Cluster Five dropped all of their courses (100% of their courses taken), even though they took on average two courses and attempted about 3 units per courses. In the last cluster, Cluster Six, students 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 the comprehension of mathematically derived clusters by the analog human brain, the study adopted an often preferred method of "naming" the clusters based on the above observations of behaviors within each cluster. By giving names to the clusters, a reader can easily associate the clusters with their demonstrated analogue characteristics.

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

An explorative research study (unsupervised data mining) like this one needs to link the results from "unsupervised" (explorative) data mining to existing fields that are familiar to people for further understanding, particularly demographics. Much useful information emerged from this exercise. The first task was to examine the differences in outcomes fields, such as GPA, persistence (defined earlier) and FTES. FTES stands for Full-time Equivalent Student, which is a productivity measure. In the state in which this study is conducted, colleges are funded by FTES.

Table 3 clearly shows the differences in the mean (μ) of the term GPA by each cluster. Some quick observations: Clusters Three (Well-adjusted Course Packers), Two (Confident Unit Loaders) had the highest term GPA and Clusters One (Careful Nibblers) and Five (Total Withdraw) the lowest.

Table 3: 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 4 shows the mean of adjustment factor, fall 2002 semester course count (CrsCnt027), and all courses ever taken by the students in the past (CrsCnt). Some quick observations: Cluster Five (Total Withdraw) contains nothing but those who completely withdrew from the college. No one in Cluster Two (Confident Unit Loaders) withdrew from any of the classes. In fall 2002, students in Cluster Three (Well-adjusted Course Packers) took on average 5 courses, while Clusters One (Careful Nibblers) and Six (Unit Maximizers) had only one. Taking their course history at the college into consideration, on average, students in Cluster Three (Well-adjusted Course Packers) still had the highest course volume, while Clusters One (Careful Nibblers) and Six (Unit Maximizers) the lowest.

Table 4: Adjustment Factor, Course Volume 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
	Count	3025	3390	4120	2205	1723	654

Table 5 shows the mean (μ) of FTES for fall 2002 and for the cumulative FTES, as well as the maximum of the number of courses taken by students. The use of maximum, in this case, helps identify the largest number of terms enrolled by the students across clusters. Some quick observations: Cluster Three (Well-adjusted Course Packers) had highest FTES both for fall 2002 and for historical cumulative. The reverse is true for Clusters One (Careful Nibblers) and Six (Unit Maximizers). However, Cluster One appeared to have the highest number of term counts, which translates into the highest number of semesters students in that cluster re-enrolled at the college.

Table 5: 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

Table 6 shows the percentage of students in fall 2002 who returned in spring 2003. Some quick observations: Clusters One (Careful Nibblers), Two (Confident Unit Loaders), Four (Overly Burdened) and Six (Unit Maximizers) had similar persistence rate of a little higher than 60%, while Cluster Three (Well-adjusted Course Packers) had close to 86% of its students returned. On the other hand, Cluster Five (Total Withdraw), those who withdrew from all courses, had the lowest rate of return (29%).

Table 6: Persistence by Clusters

	TwoStep Clusters											
	CL-1		CL-2		CL-3		CL-4		CL-5		CL-6	
	RETURN		RETURN		RETURN		RETURN		RETURN		RETURN	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
0	1149	38.0%	1323	39.0%	589	14.3%	751	34.1%	1224	71.0%	240	36.7%
1	1876	62.0%	2067	61.0%	3531	85.7%	1454	65.9%	499	29.0%	414	63.3%

The following tables contain the analysis of rank-ordered outcomes fields by clusters. These fields are Term GPA, FTES for fall 2002, Persistence, cumulative FTES (FTES History), and term counts (Attendance History). The values (means) have been ranked from 1 to 6 with 6 being the highest. To convert raw values into rank ordered values helps with standardizing the scales of performance to the extent possible. Note the tables use the names of the clusters.

The Well-adjusted Course Packers (Cluster Three) in Table 7 had the highest performance, while the Total Withdraws (Cluster Five) had the lowest. Other clusters fell somewhere in between.

Table 7: Clusters by Term GPA, Term FTES, and Persistence

		Term		
Clusters	**Name**	**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

Table 8 below brought in some historical perspective by adding cumulative FTES and attendance history. It seems that even though the Total Withdraw (Cluster Five) were low in generating FTES for the term they were studied (Table 7) but they certainly have produced high FTES for the college over the years (Table 8). The Overly Burdened (Cluster Four) also produced high FTES although they appeared to be struggling in fall 2002. Those Careful Nibblers (Cluster One) generated very little FTES, but they remained enrolled/engaged with the college throughout the years.

Table 8: Clusters by Term GPA, FTES History and Attendance History

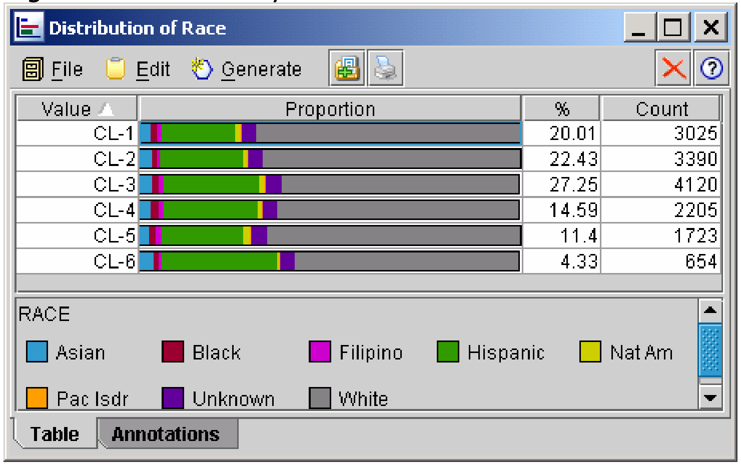
Clusters	Name	Term GPA	FTES History	Attendance History
1	Careful Nibblers	2	1	5
2	Confident Unit Loaders	5	3	6
3	Well-adjusted Course Packers	6	6	4
4	Overly Burdened	3	5	3
5	Total Withdraw	1	4	2
6	Unit Maximizers	4	2	1

The natural progression of explorative research into the clusters would prompt one to query the distribution of demographic fields within each cluster.

Figures 4 through 8 are proportion graphs to display the distribution of race and enrollment status of the students by each cluster. Visually speaking, in Figure 4, the different categories of race are close to evenly distributed. Figure 5 shows even distribution (or close to) of different enrollment status of the students, with Well-adjusted Course Packers (Cluster Three) and Overly Burdened (Cluster Four) showing more first-time students. Interestingly, these two clusters are similar in term GPA, but different in FTES and Persistence. Except for the "X" category of gender, which means "Unknown/Unreported", gender is also evenly distributed across clusters. Compared to other proportion graphs, race, gender, and enrollment status do not warrant, at this point, as much attention as student age and educational goal.

With typological groupings based on behaviors, every subject is equal, but also special in the study. The subjects are treated equally because they all contribute to the behaviors. Their behaviors are then treated specially in a Euclidean Hyperspace. Using demographics in creating clusters would bring in factors out of the control of the students and probably the college as well. On the other hand, using the demographics to examine the clusters would provide far meaningful information. It is almost like giving personalities to the clusters.

Figure 4: Clusters by Race



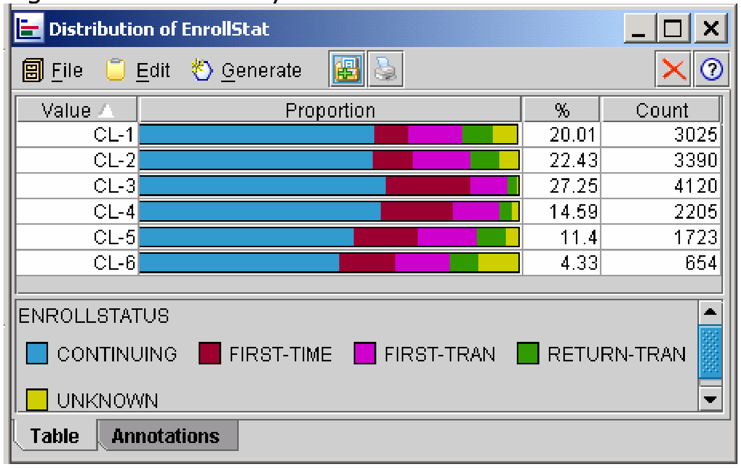
Value	Proportion	%	Count
CL-1		20.01	3025
CL-2		22.43	3390
CL-3		27.25	4120
CL-4		14.59	2205
CL-5		11.4	1723
CL-6		4.33	654

RACE

🟦 Asian 🟥 Black 🟪 Filipino 🟩 Hispanic 🟨 Nat Am

🟧 Pac Isdr 🟪 Unknown ⬛ White

Figure 5: Clusters by Enrollment Status



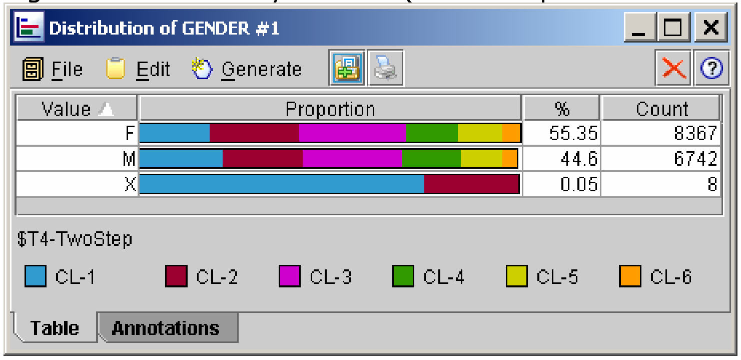
Value	Proportion	%	Count
CL-1		20.01	3025
CL-2		22.43	3390
CL-3		27.25	4120
CL-4		14.59	2205
CL-5		11.4	1723
CL-6		4.33	654

ENROLLSTATUS

🟦 CONTINUING 🟥 FIRST-TIME 🟪 FIRST-TRAN 🟩 RETURN-TRAN

🟨 UNKNOWN

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

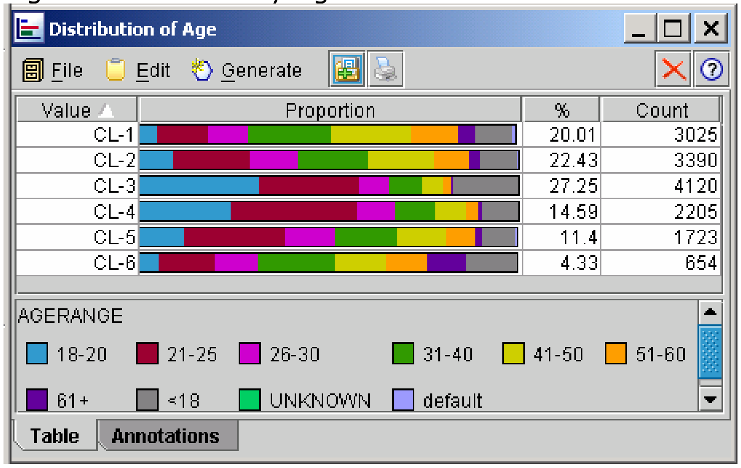


Value	Proportion	%	Count
F		55.35	8367
M		44.6	6742
X		0.05	8

$T4-TwoStep

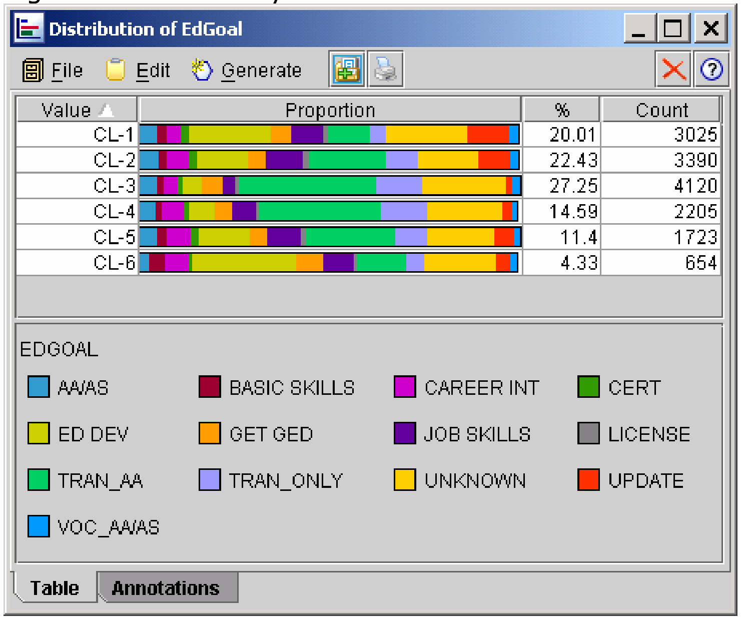
🟦 CL-1 🟥 CL-2 🟪 CL-3 🟩 CL-4 🟨 CL-5 🟧 CL-6

Figure 7: Clusters by Age



Value	Proportion	%	Count
CL-1		20.01	3025
CL-2		22.43	3390
CL-3		27.25	4120
CL-4		14.59	2205
CL-5		11.4	1723
CL-6		4.33	654

Figure 8: Clusters by Educational Goals



Value	Proportion	%	Count
CL-1		20.01	3025
CL-2		22.43	3390
CL-3		27.25	4120
CL-4		14.59	2205
CL-5		11.4	1723
CL-6		4.33	654

EDGOAL

🟦 AA/AS	🟥 BASIC SKILLS	🟪 CAREER INT	🟩 CERT
🟨 ED DEV	🟧 GET GED	🟪 JOB SKILLS	⬛ LICENSE
🟩 TRAN_AA	🟪 TRAN_ONLY	🟨 UNKNOWN	🟥 UPDATE
🟦 VOC_AA/AS			

Graphic display of student age and educational goal by cluster showed distinct differences that prompted the study author to further examine them using Chi-square analysis. Prior to this analysis, the study generated a web graph using age and clusters to visually confirm the differences in age by clusters.

Figure 9: Web Graph of Age and Clusters

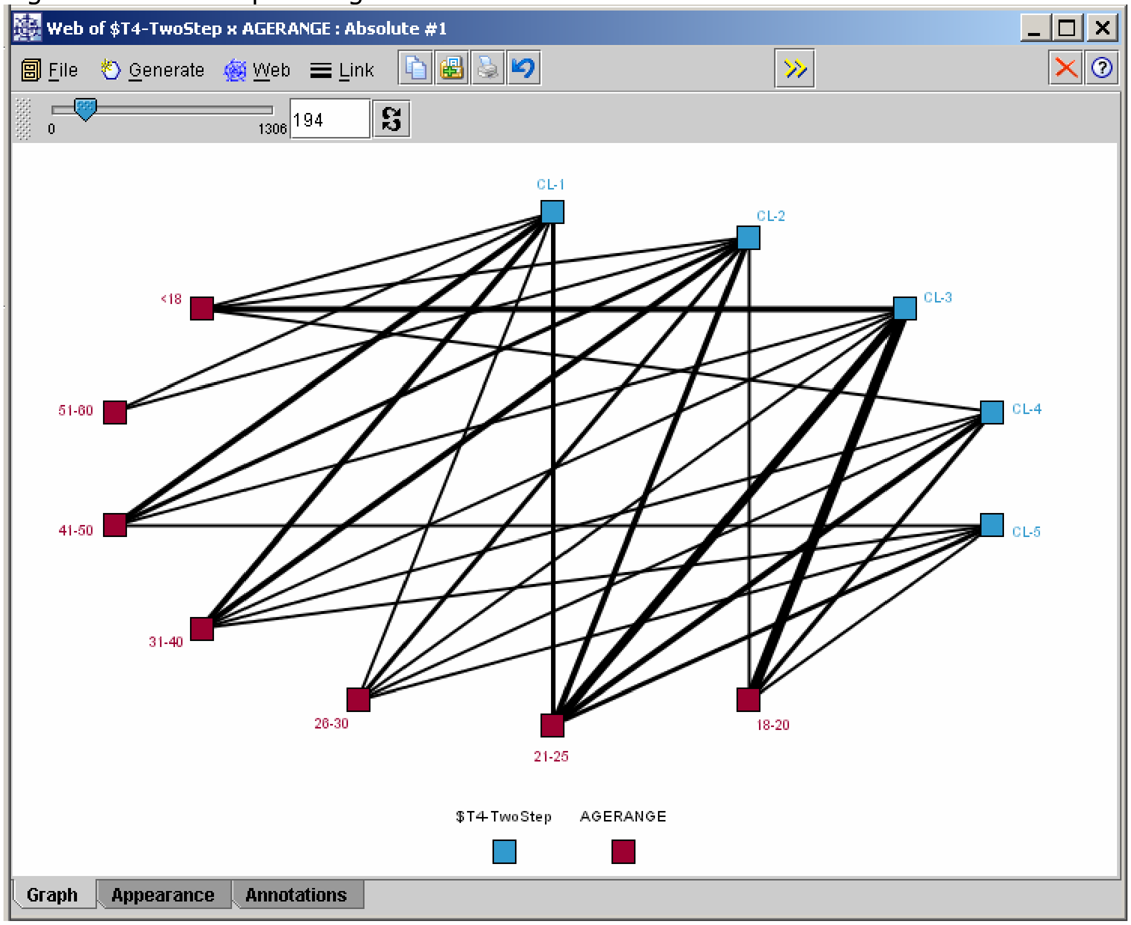


Figure 9 is a special data graph called "Web" graph, which helps indicate the strength of associations between fields (variables). In this case, it is clearly visible via the thicker lines the strong relationship between students aged 18-20 and their presence in Cluster Three (Well-adjusted Course Packers).

Table 9 presents the 10 age groupings (AgeRank) by clusters with observed and expected observations. The differences are significant at .0001 level based on Asymptotic 2-sided (*X*(45), p< .0001). For the cross-tabulation between educational goals and clusters, 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 is necessary.

Clusters Three and Four had similar high numbers of students aged 18-20, and students with transfer goals. Even though the term GPA indicated Clusters Two and Three are similar, but the demographic analysis seems to indicate Clusters Three and Four are similar.

Table 9: Chi-square Cross Tabulation of Age by Clusters

AGERANGE * $T4-TwoStep Crosstabulation

			$T4-TwoStep						
			CL-1	CL-2	CL-3	CL-4	CL-5	CL-6	Total
AGERANGE	<18	Count	298	323	715	216	153	88	1793
		Expected Count	358.8	402.1	488.7	261.5	204.4	77.6	1793.0
		% within $T4-TwoStep	9.9%	9.5%	17.4%	9.8%	8.9%	13.5%	11.9%
	18-20	Count	144	302	1307	531	204	34	2522
		Expected Count	504.7	565.6	687.3	367.9	287.5	109.1	2522.0
		% within $T4-TwoStep	4.8%	8.9%	31.7%	24.1%	11.8%	5.2%	16.7%
	21-25	Count	408	688	1071	736	461	97	3461
		Expected Count	692.6	776.1	943.3	504.8	394.5	149.7	3461.0
		% within $T4-TwoStep	13.5%	20.3%	26.0%	33.4%	26.8%	14.8%	22.9%
	26-30	Count	322	432	338	220	226	74	1612
		Expected Count	322.6	361.5	439.3	235.1	183.7	69.7	1612.0
		% within $T4-TwoStep	10.6%	12.7%	8.2%	10.0%	13.1%	11.3%	10.7%
	31-40	Count	668	624	368	228	287	133	2308
		Expected Count	461.8	517.6	629.0	336.7	263.1	99.8	2308.0
		% within $T4-TwoStep	22.1%	18.4%	8.9%	10.3%	16.7%	20.3%	15.3%
	41-50	Count	637	584	220	178	224	89	1932
		Expected Count	386.6	433.3	526.5	281.8	220.2	83.6	1932.0
		% within $T4-TwoStep	21.1%	17.2%	5.3%	8.1%	13.0%	13.6%	12.8%
	51-60	Count	376	320	81	76	133	73	1059
		Expected Count	211.9	237.5	288.6	154.5	120.7	45.8	1059.0
		% within $T4-TwoStep	12.4%	9.4%	2.0%	3.4%	7.7%	11.2%	7.0%
	61+	Count	146	101	15	19	30	65	376
		Expected Count	75.2	84.3	102.5	54.8	42.9	16.3	376.0
		% within $T4-TwoStep	4.8%	3.0%	.4%	.9%	1.7%	9.9%	2.5%
	default	Count	21	16	4	1	5	1	48
		Expected Count	9.6	10.8	13.1	7.0	5.5	2.1	48.0
		% within $T4-TwoStep	.7%	.5%	.1%	.0%	.3%	.2%	.3%
	UNKNOWN	Count	5	0	1	0	0	0	6
		Expected Count	1.2	1.3	1.6	.9	.7	.3	6.0
		% within $T4-TwoStep	.2%	.0%	.0%	.0%	.0%	.0%	.0%
Total		Count	3025	3390	4120	2205	1723	654	15117
		Expected Count	3025.0	3390.0	4120.0	2205.0	1723.0	654.0	15117.0
		% within $T4-TwoStep	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
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 10: Chi-square Cross Tabulation of Educational Goals by Clusters

EDGOAL * $T4-TwoStep Crosstabulation

			$T4-TwoStep						
			CL-1	CL-2	CL-3	CL-4	CL-5	CL-6	Total
EDGOAL	AA/AS	Count	135	170	188	96	79	16	684
		Expected Count	136.9	153.4	186.4	99.8	78.0	29.6	684.0
		% within $T4-TwoStep	4.5%	5.0%	4.6%	4.4%	4.6%	2.4%	4.5%
	BASIC SK	Count	76	65	62	38	40	27	308
		Expected Count	61.6	69.1	83.9	44.9	35.1	13.3	308.0
		% within $T4-TwoStep	2.5%	1.9%	1.5%	1.7%	2.3%	4.1%	2.0%
	CAREER I	Count	120	203	151	129	106	42	751
		Expected Count	150.3	168.4	204.7	109.5	85.6	32.5	751.0
		% within $T4-TwoStep	4.0%	6.0%	3.7%	5.9%	6.2%	6.4%	5.0%
	CERT	Count	66	72	47	27	37	6	255
		Expected Count	51.0	57.2	69.5	37.2	29.1	11.0	255.0
		% within $T4-TwoStep	2.2%	2.1%	1.1%	1.2%	2.1%	.9%	1.7%
	ED DEV	Count	657	460	201	149	230	179	1876
		Expected Count	375.4	420.7	511.3	273.6	213.8	81.2	1876.0
		% within $T4-TwoStep	21.7%	13.6%	4.9%	6.8%	13.3%	27.4%	12.4%
	GET GED	Count	161	163	228	105	77	47	781
		Expected Count	156.3	175.1	212.9	113.9	89.0	33.8	781.0
		% within $T4-TwoStep	5.3%	4.8%	5.5%	4.8%	4.5%	7.2%	5.2%
	JOB SKIL	Count	250	324	147	138	155	52	1066
		Expected Count	213.3	239.1	290.5	155.5	121.5	46.1	1066.0
		% within $T4-TwoStep	8.3%	9.6%	3.6%	6.3%	9.0%	8.0%	7.1%
	LICENSE	Count	44	59	39	20	25	6	193
		Expected Count	38.6	43.3	52.6	28.2	22.0	8.3	193.0
		% within $T4-TwoStep	1.5%	1.7%	.9%	.9%	1.5%	.9%	1.3%
	TRAN_AA	Count	327	686	1490	706	409	86	3704
		Expected Count	741.2	830.6	1009.5	540.3	422.2	160.2	3704.0
		% within $T4-TwoStep	10.8%	20.2%	36.2%	32.0%	23.7%	13.1%	24.5%
	TRAN_ONL	Count	131	279	510	271	146	31	1368
		Expected Count	273.7	306.8	372.8	199.5	155.9	59.2	1368.0
		% within $T4-TwoStep	4.3%	8.2%	12.4%	12.3%	8.5%	4.7%	9.0%
	UNKNOWN	Count	650	547	897	434	307	125	2960
		Expected Count	592.3	663.8	806.7	431.8	337.4	128.1	2960.0
		% within $T4-TwoStep	21.5%	16.1%	21.8%	19.7%	17.8%	19.1%	19.6%
	UPDATE	Count	327	290	76	60	85	26	864
		Expected Count	172.9	193.8	235.5	126.0	98.5	37.4	864.0
		% within $T4-TwoStep	10.8%	8.6%	1.8%	2.7%	4.9%	4.0%	5.7%
	VOC_AA/A	Count	81	72	84	32	27	11	307
		Expected Count	61.4	68.8	83.7	44.8	35.0	13.3	307.0
		% within $T4-TwoStep	2.7%	2.1%	2.0%	1.5%	1.6%	1.7%	2.0%
Total		Count	3025	3390	4120	2205	1723	654	15117
		Expected Count	3025.0	3390.0	4120.0	2205.0	1723.0	654.0	15117.0
		% within $T4-TwoStep	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

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.

SUMMARY OF DISCOVERIES

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

- Careful Nibblers (Cluster One) were low on every aspect, except for being continually enrolled at the college for many years (having high historical attendance).
- Neither Careful Nibblers (Cluster One) nor Confident Unit Loaders (Cluster Two) produced high FTES, but they maintain a long history of attending the college.
- Confident Unit Loaders (Cluster Two) differed from Careful Nibblers (Cluster One) by having high term GPA.
- Well-adjusted Course Packers (Cluster Three) had been 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 (Cluster Three) and Overly Burdened (Cluster Four) had more traditional transfer directed students.
- Overly Burdened (Cluster Four) appeared to be students who were performing less well as Well-adjusted Course Packers (Cluster Three), even though they are far more alike in demographics compared to other clusters.
- Total Withdraw (Cluster Five), 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 (Cluster Six) were similar to Careful Nibblers in demographics. They were mediocre 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, failed to demonstrate their ability to account for much of the variance across clusters.

DISCUSSIONS

These typologies have a number of serious and practical implications. Careful Nibblers are likely lifelong learning students who live and work within the service community, therefore, 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. The same can be said about the second cluster, the Confident Unit Loaders. There are more younger students present in this cluster (Figure 7), who can be motivated to become transfer students if their needs are carefully studied.

The third cluster, Well-adjusted Course Packers, is a positive contributing factor for a variety of college performance outcomes, such as transfer, persistence, success, even time to degree. They are younger students in majority and 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 well. It is important to study this cluster closely because they have the potential of becoming the stars in Cluster Three (Well-adjusted Course Packers).

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

Unit Maximizers are a group of students who demonstrated a fly-by-night 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 shot stay makes them less easy to work with. How to attract them to stay? Many of them have similar goal of "Educational Development" (Figure 7), as Careful Nibblers do, so perhaps they can move beyond this goal into something higher.

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 have been overused in predicting students' outcomes. They will remain important elements to report out college data, yet, there is increasing practical difficulty to use them due to increased reluctance of students in stating their biological stats.

Further analysis beyond the scope of this study may be the following:

1) To drill further into the dynamics of clusters by studying the distribution of students by academic departments by clusters.
2) Evaluate the use of other behavior related fields for developing clusters.
3) Conduct predictive modeling by clusters to determine the accuracy of predicting students GPA, Persistence, FTES.
4) Classify future students into the 6 clusters (or whatever final scenarios may be) and conduct GPA, FTES, and Persistence projections for the students.

CONCLUSION

This explorative data mining project used a less visited statistical analysis technique, distance based clustering algorithm, to study 3 sub-indices of the OIndex and stabilized at a 6-cluster scenario following an exhaustive explorative study of 4, 5, and 6 cluster scenarios produced by K-Means and TwoStep algorithms.

Using principles in data mining, the study followed a proven data mining process that proceeded from identifying the research questions, to staging the data, to data auditing, and to building scenarios. All scenarios were subjected to data visualization, and in cases appropriate, Chi-square analysis.

This study established 6 typologies of students enrolled at a suburban community college. It based the study on the notion that indicators, called indexes in the study, of student behavioral data are good candidates for new facets of research studies, compared to using non-behavioral data, such as gender or race. Information from behavioral indexes, called OIndex in this study, emerged as both meaningful for understanding and measuring students' learning as well as actionable for decision making. The typologies may be added to existing educational strategies for both management and assessment of learning.

Bibliography

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.

Fenske, R., Keller, J., & Irwin, G. (1999). Toward a Typology of Early Intervention Programs. Advances in Education Research. Vol. 4.

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

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

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.

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