Multi-Dimensional Representations -- How Many Dimensions?

Cluster Stack Visualization for Market Segmentation Analysis

William Wright
Oculus Info Inc.

An earlier version of this paper appeared in the
Symposium on New Paradigms in Information Visualization and Manipulation,
ACM Conference on Information and Knowledge Management, November, 1997.

1. Introduction

Information visualization methods present multi-dimensional data graphically by mapping data and properties to different visual shapes, colors, and positions [3]. Visualizations can be in a 2-D or 3-D Euclidean space but often portray many more dimensions. For example, six dimensions can easily be shown by a cube's x, y and z dimension, the cube's color, and it's position on an x, z plane.

The purpose of visualization is to improve comprehension and reveal significance. It is easier to comprehend a visual representation than a numerical one, especially when the visual representation replaces many pages or screens of data.

Interaction methods, such as motion, navigation, filtering, drill down brushing [1], selection and comparison, increase the usability of a visualization, but an appropriate visual design is the crucial first step to creating an effective visualization. As Edward Tufte [2] points out, good visual designs provide:

- Complex ideas communicated with clarity, precision and efficiency;
- Present many things in a small space;
- Encourage data comparison;
- Reveal data at several levels of detail; and
- Induce the viewer to think about the substance.

Visual design is important for high dimensionality visualizations, and Tufte's mandates may be used as guidance and evaluation criteria.

2. How Many Dimensions? 2-D, 3-D, … or n-D

This paper describes a new visualization technique for very high dimensional domains, and discusses the technique's strengths and weaknesses. This new technique, called Cluster Stacks, seems to be appropriate for visualizing 100 to 200 or more dimensions, concurrently.

This new 3-D visualization technique also helps illustrate the role of 3-D vs. 2-D. Graphical displays of information in 2-D seem to be more suitable when there are smaller amounts of data, and/or a smaller number of dimensions to view, and when the task being supported is focused with little if any context required. Otherwise, when there are larger amounts of data, and/or many dimensions to view, 3-D seems to be a more suitable display. Further, when we say 3-D, we are really referring to an Euclidean space. Often in information visualization, we are really viewing n-D properties of objects. In the case presented here, we are viewing 158 dimensions.

This paper also shows how information visualization complements data mining analytical techniques by visualizing the output of those techniques. Both data mining and information visualization technologies are directed at gaining business value from vast amounts of under-utilized business data.

3. The Application -- Segmentation and Clustering

Data Mining tools help uncover meaningful correlations, patterns and trends in large amounts of data through statistical pattern recognition techniques. Often termed Knowledge Discovery in Databases (KDD), it is the extraction of previously unknown information from data, and encompasses a number of techniques including classification, association rules, decision trees, neural networks, and clustering.

Clustering is the partitioning of a data universe into sets so that all members of each set are similar according to some metric. The members are grouped together because of their similarity or proximity across a variety of characteristics. Numerical fields for characteristics are normalized (e.g. have a range between 0 and 1), and typically a multi-component Euclidean distance metric is used to compute distance between clusters. (Clustering algorithms are available in a number of commercial statistical tools.) Clustering easily handles large numbers of observations with many dimensions. The example shown here has approximately 2.3 million observations and 158 dimensions.

While clustering is the name of one analytical technique, segmentation refers to the general practice of the marketing profession to decompose the market for a product or service into several distinct types or segments of customers. These distinctly different sub-populations differ in product utilization, purchasing habits, lifestyles and demographics for example. Developing a market requires an accurate assessment of these segments in order to optimize product communication and delivery. In a marketing organization, the objective is to hit the targets, once defined, and be able to concentrate all elements of the marketing mix against the target.

Many companies use statistical analysis of their large and comprehensive customer/product databases of individual customer behaviors (e.g. when and how products are purchased) to generate descriptions of their segments.

In practice, the use of clustering poses a number of difficulties due to the complexity of the data and the process. Running a statistical clustering algorithm on a massive customer/product database yields clusters. But each cluster is described by 100 to 200 different behaviors, or dimensions, and their differences are complex. Expert marketers understand the resulting 10 to 30 clusters, but only after spending days pouring over spreadsheet after spreadsheet of results. When this information is visualized, even the experts are surprised by the bad data and the unexpected relationships. But the expert marketers need to be able to turn around and communicate the rich complex story within the clusters to the business managers who make product and customer decisions. Rows and columns of numbers are ineffective. Nicknames (e.g. “the recently retired” or “soccer moms”) are offensive and really are a univariate description. This situation is an opportunity to use visualization as a tool to help understand and present complex data and relationships.

To be relevant and useful, the visualization needs to support the business objectives, which are to:

- Understand customer behavior; and the richness and variability of the behaviors within clusters;
- Look at all of the data simultaneously both in aggregate, and by segment in aggregate; and
- Facilitate understanding and communication to senior management of the segment characteristics, especially when the differences between clusters are subtle;

4. The Visualization

The objective of this high-dimensional “Cluster Stack” system is to provide a visual description of the behaviors within a cluster. Users need to be able to easily see who are the customers, how do they differ, and what are their major characteristics. The Cluster Stack visualization should reveal, and make it possible to use, the full richness and variability of the cluster characteristics.

Two types of activities are supported by the system: analysis and communication. An analyst uses the

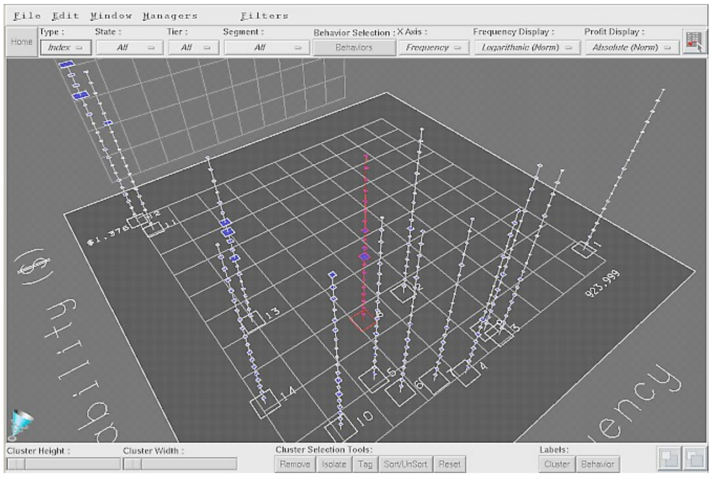
Cluster Stack visualization during the development of the clustering analysis to review early results, and to check on data quality, and the operation of the clustering algorithms. After the final cluster computation is done, the analyst makes observations about clusters and behaviors, and then communicates the findings to the business managers responsible for different products and market segments. During meetings with the business managers, the visualization is used as a discussion vehicle to explore the opportunities for changing behaviors.

Figure 1 shows the layout of the Cluster Stack visualization. The ground plane is a scatter plot for the clusters. Each cluster is identified by a number next to it in the ground plane. Along the front is the "Frequency" axis which shows the population size of each cluster. On the left is the "Profitability" axis. The height shows behavioral dimensions for the clusters. Each cluster is represented as a multi-level tree. Each level of the tree shows a behavior as a rectangle. The score's value for that behavior for that cluster is mapped to both the width and depth of the rectangle.

The top menu bar allows selection of a data set (e.g. by geography, by tier), the selection of behaviors (see the Dialog Box in Figure 4), and the selection of type of score for the behaviors. Three scores are possible: 1) Index which indicates how much a behavior serves to differentiate a cluster from other clusters, 2) Rank which indicates a behavior's overall rank as a differentiating behavior, and 3) Mean of the behavior's value for a cluster. All figures in this paper show Index because it most clearly shows how clusters differ.

Behaviors are grouped by Product. A list of Products is shown in the Dialog Box in Figure 4. In this visualization, there are 158 behaviors in total.

Figures 1 through 7 show different behaviors of interest for 14 clusters. Note that the data set has been adjusted to protect confidentiality. The Figures are screen shots from the operational system, and it should be noted that the system provides full motion and interaction on an SGI O2 R5000 workstation. Almost similar performance would be expected on a Pentium Pro.

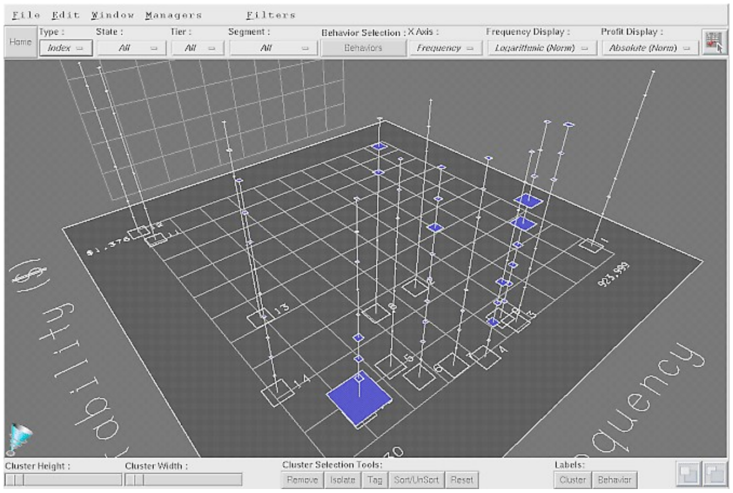


Type : Index
State : All
Tier : All
Segment : All
Behavior Selection : Behaviors
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

Figure 1 - Basic Demographics
Here we can see what demographic and product factors differentiate the clusters. For example, Cluster 13 has larger Retirement and Investment balances. Cluster 8 has a larger Savings balance. Many clusters appear equally thin (i.e. are not differentiated by these factors). Factor weights are shown as rectangles top-to-bottom in each tree. The factors shown in Figure 1, top-to-bottom, are listed in Table 1.

Bank Card:	Amount of Purchases
Mortgage:	Orig. Amount Balance
Line of Credit:	Balance
Retail Loan:	Balance Credit Score
Retirement:	Balance
Investment:	Balance
Cert. of Dep.:	Balance
Money Market:	Balance
Savings:	Balance
Checking:	Number of Debits
Household:	Length of Time of Resid. Liklihood of Being a Professional Liklihood of Being Married Income Home Value # of Children Age

Table 1. Basic Demographics. Total of 19 dimensions. Listed as shown top to bottom in Figure 1.



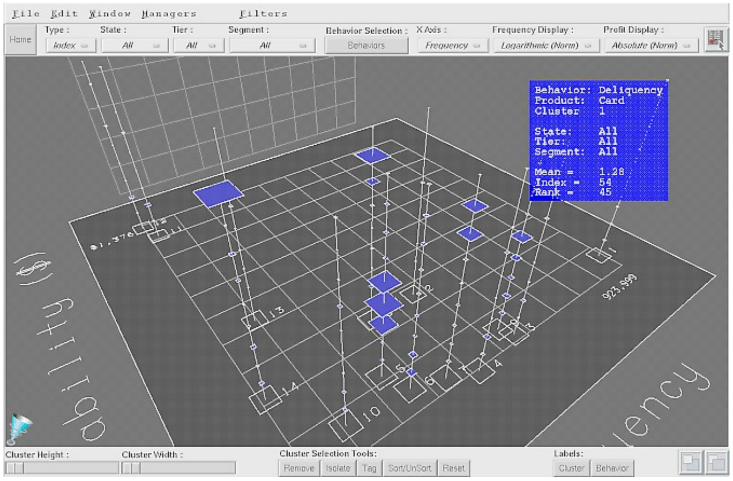
Type : Index
State : All
Tier : All
Segment : All
Behavior Selection :
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

Figure 2 - ATMs vs. Tellers
Considering just the Checking product, here we can see if the ATM or Teller channel is important. Clusters 8, 3 and 6 use tellers. Clusters 10, 4 and 9 use ATMs. Cluster 9 uses foreign ATMs. ATM vs. Teller utilization is not an important differentiating factor for the remaining clusters.

Checking:

Teller Inquiry Freq.
Teller Debit Freq.
Teller Credit Freq.
Foreign ATM Inquiry Freq.
Foreign ATM Debit Freq.
# of Unique ATM Transaction Dates
ATM Inquiry Freq.
ATM Debit Freq.
ATM Credit Freq.

Table 2. ATM vs. Tellers.
Total of nine dimensions. Factor weights are shown as rectangles top-to-bottom in each tree in Figure 2. The factors, top-to-bottom, are listed above in Table 2.



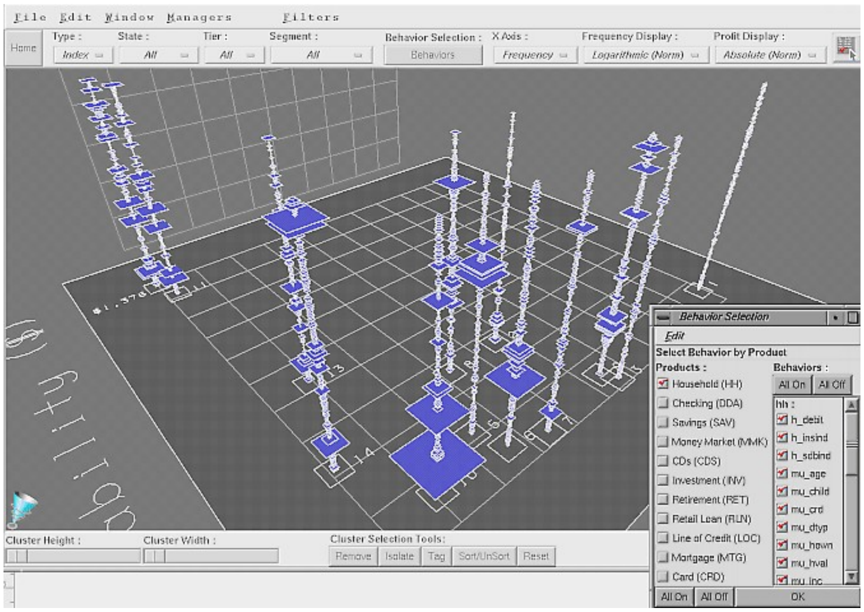
Type : Index
State : All
Tier : All
Segment : All
Behavior Selection :
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

```
Behavior: Deliquency
Product:  Card
Cluster   1

State:    All
Tier:     All
Segment:  All

Mean =    1.28
Index =   54
Rank =    45
```

Figure 3 - Credit Dimensions
Here we can see the importance of credit factors. Cluster 5 appears to be in the habit of making late payments on their Retail Loans. Cluster 14 has a problem with Bank Cards. The blue text is a mouse-over drill down display of details of one factor or cluster.

Bank Card:
Delinquency
Line of Credit:
# Times Past Due 60+ Days
# Times Past Due 30 Days
Credit Score
Retail Loan:
# Times Past Due 90 Days
# Times Past Due 60 Days
# Times Past Due 30 Days
Money Market:
NSF Freq.
Checking:
NSF Freq.

Table 3. Credit.
Total of nine dimensions. Factor weights are shown as rectangles top-to-bottom in each tree in Figure 3. The factors, top-to-bottom, are listed above in Table 3.



Type : Index
State : All
Tier : All
Segment : All
Behavior Selection : Behaviors
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

Behavior Selection

Edit

Select Behavior by Product

Products :

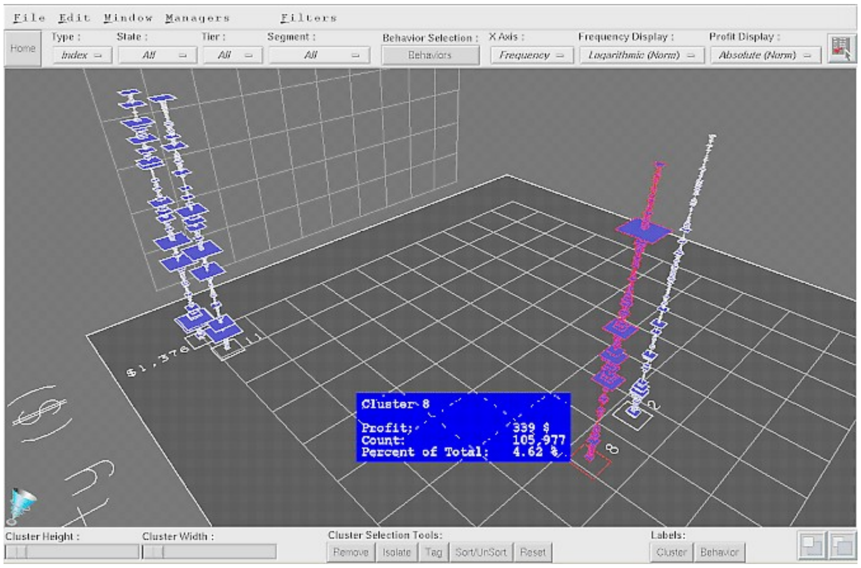
☑ Household (HH)
☐ Checking (DDA)
☐ Savings (SAV)
☐ Money Market (MMK)
☐ CDs (CDS)
☐ Investment (INV)
☐ Retirement (RET)
☐ Retail Loan (RLN)
☐ Line of Credit (LOC)
☐ Mortgage (MTG)
☐ Card (CRD)

Behaviors :

hh :
☑ h_debit
☑ h_insind
☑ h_sdbind
☑ mu_age
☑ mu_child
☑ mu_crd
☑ mu_dtyp
☑ mu_hown
☑ mu_hval
☑ mu_inc

Figure 4 - All 158 Dimensions

The Cluster Stack visualization can easily show many dimensions concurrently. Key differentiating behaviors can be identified and compared across clusters. It is interesting to note that Cluster 1, on the right and the cluster with the largest population, does not have any differentiating behaviors, and perhaps can be considered the "moderate majority".

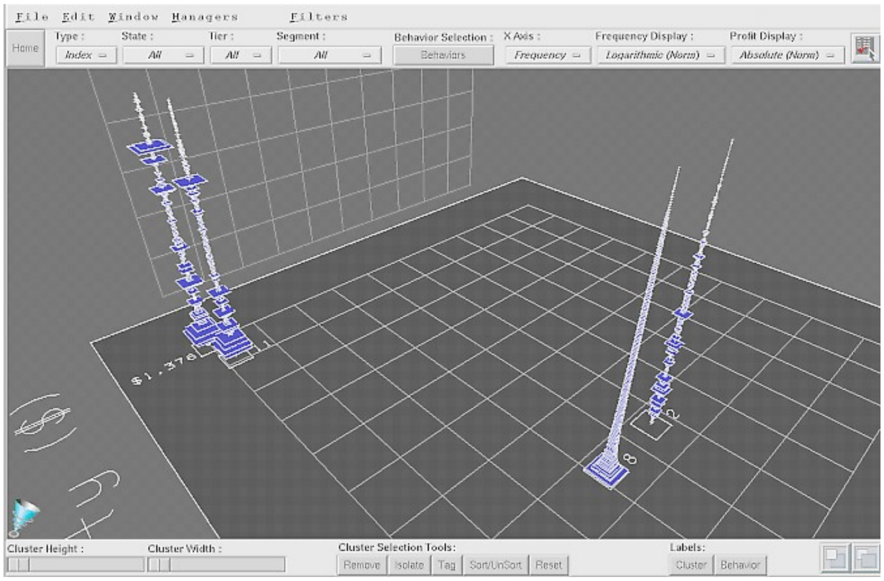


Type : Index
State : All
Tier : All
Segment : All
Behavior Selection : Behaviors
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

```
Cluster 8

Profit;            339 $
Count:             105,977
Percent of Total:  4.62 %
```

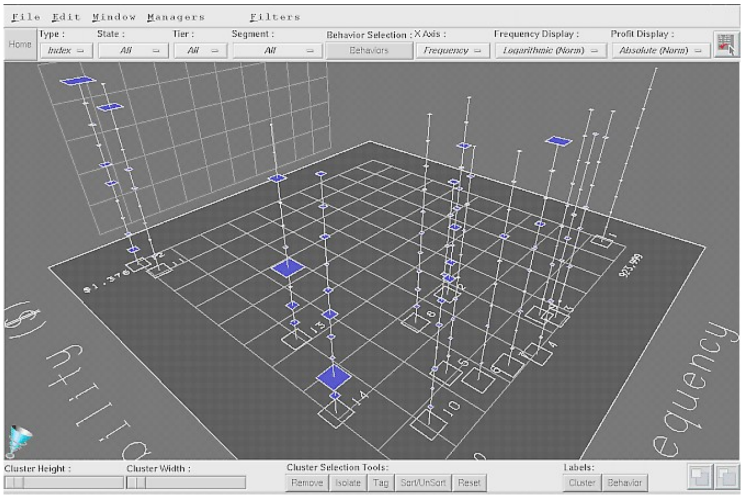
Figure 5 - Comparing Clusters

Clusters can be toggled on and off to allow users to focus on, and compare, selected clusters. Here the two most profitable clusters, on the left, are compared with two moderately profitable clusters with larger populations.



Type : Index
State : All
Tier : All
Segment : All
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

Figure 6 - Comparing Sorted Dimensions
To help compare one cluster against others, a cluster's behaviors, in this case Cluster 8, can be sorted. The new ordering of behaviors is applied to all Clusters.



Type : Index
State : All
Tier : All
Segment : All
X Axis : Frequency
Frequency Display : Logarithmic (Norm)
Profit Display : Absolute (Norm)

Figure 7 - Product Penetration
As a final picture of customers, Figure 7 shows product utilization by cluster.

Products

Mortgage
Retail Loan
Retirement
Savings
Line of Credit
Bank Card
Cert of Deposit
Checking
Money Market
Investment

Table 4. Product Penetration. Total of ten dimensions. Factor weights are shown as rectangles top-to-bottom in each tree in Figure 7. The factors, top-to-bottom, are listed above in Table 4.

The viewer's point-of-view of the visualization can be turned and rotated, this way and that, as the user examines the data, the clusters and their relationships. Also of note in the Cluster Stack visualization, is the use of drill-down brushing (i.e. the ability to point at any object, clusters or behaviors in this case, and display precise numerical / text descriptions).

5. Conclusions

This visualization system, as described above, is in use by a marketing analysis group within a large financial institution. The visualization allows users to do "quick visual arithmetic" and make significant observations and comparisons, with the details always at hand. A number of improvements and extensions have been identified. However, considering just the visualization aspects, the following strengths and weaknesses have been noted.

Strengths:

- The core visual paradigm is clean , simple and very readable. It is easy to learn, and easy to use. Clusters are described in detail without overwhelming the users.
- Many dimensions can be displayed, and the selection of dimensions, whether all of the dimensions, or a subset of interest, is useful because it allows different issues to be explored.
- A user workflow of "context and focus, context and focus" is supported.
- The comparison of clusters, their differences and similarities, is quick and efficient.
- Occlusion, if it occurs, can be easily removed by spinning the display until all clusters are clearly in view, or by interactively scaling one or two dimensions to increase the separation among closely grouped clusters.
- The visual paradigm allows extensions. For instance, more than one behavior attribute can be mapped to a rectangle.

Weaknesses:

- Identification of the behaviors within the clusters, while provided now via brushing, is not as easy as it could be. Too much user effort is spent identifying behaviors rather than interpreting their significance. Selective behavior labels within the landscape would be useful. Only the behaviors of interest should be labeled.
- Similarly, product identification could be better supported. Each group of behaviors associated with a product could be color coded.
- Comparison of individual behaviors across clusters is difficult when all 158 behaviors are displayed. Allowing the user to highlight several behaviors across all clusters would facilitate an understanding of those behaviors within the context of all behaviors.

The Cluster Stack visualization is a new kind of information visualization application and is illustrative of an emerging trend to see visualization tools coupled with data mining tools to enhance the comprehension and interpretation of the data mining results.

It is, in a way, extraordinary to think it is possible to see 158 dimensions at once on a single screen. A good (according to Tufte) visualization design makes it possible. While 3-D visualization is a powerful medium, perhaps it should be considered n-D instead of 3-D.

6. References

[1] Cleveland, W.S. and M.E. McGill, Dynamic Graphics for Statistics, Wadsworth, Belmont, Calif., 1988.

[2] Tufte, E.R., The Visual Display of Quantitative Information, Graphics Press, 1983.

[3] Wright, W., Business Visualization Applications, IEEE Computer Graphics and Applications, July/August, 1997, Volume 17, Number 14.