3D Grand Tour for Multidimensional Data and Clusters

Li Yang

Institute of High Performance Computing, National University of Singapore
89 Science Park Drive, #01-05/08 The Rutherford, Singapore 118261
yangli@ihpc.nus.edu.sg

Abstract. Grand tour is a method for viewing multidimensional data via linear projections onto a sequence of two dimensional subspaces and then moving continuously from one projection to the next. This paper extends the method to 3D grand tour where projections are made onto three dimensional subspaces. 3D cluster-guided tour is proposed where sequences of projections are determined by cluster centroids. Cluster-guided tour makes inter-cluster distance-preserving projections under which clusters are displayed as separate as possible. Various add-on features, such as projecting variable vectors together with data points, interactive picking and drill down, and cluster similarity graphs, help further the understanding of data. A CAVE virtual reality environment is at our disposal for 3D immersive display. This approach of multidimensional visualization provides a natural metaphor to visualize clustering results and data at hand by mapping the data onto a time-indexed family of 3D natural projections suitable for human eye's exploration.

1 Introduction

Visualization techniques have proven to be of high value in exploratory data analysis and data mining. For data with a few dimensions, scatterplot is an excellent means for visualization. Patterns could be efficiently unveiled by simply drawing each data point as a geometric object in the space determined by one, two or three numeric variables of the data, while its size, shape, color and texture determined by other variables of the data. The ability to draw scatterplots is a common feature of many visualization systems. Conventional scatterplots lose their effectiveness, however, as dimensionality of data becomes large.

An idea comes out, then, to project higher dimensional data orthogonally onto lower dimensional subspaces. It allows us to look at multidimensional data in a geometry that is within the perceptibility of human eyes. Since there is an infinite number of possibilities to project high dimensional data onto lower dimensions, and information will eventually lose after the projection, the grand tour[1,3] and other projection pursuit techniques[10,12] aim at automatically finding the interesting projections or at least helping the users to find them.

Grand tour is an extension of data rotation for multidimensional data sets. It is based on selecting a sequence of linear projections and moving continuously from one projection to the next. By displaying a number of intermediate

D.J. Hand, J.N. Kok, M.R. Berthold (Eds.): IDA'99, LNCS 1642, pp. 173–184, 1999.

projections obtained by interpolation, the entire process creates an illusion of continuous, smooth motion through multidimensional displays. This helps to find interesting projections which is hard to find in the original data, owing to the curse of dimensionality. Furthermore, grand tour allows viewers to easily keep track of a specific group of data points throughout a tour. By examining where the data points go from one projection to the next, viewers have a much better understanding about data than using conventional visualization techniques such as bar charts or pie charts.

Now the question becomes how to choose "meaningful" projections and projection sequences to maximize the chance of finding interesting patterns. One simple way is choosing the span of any three arbitrary variables as a 3D subspace and then moving from this span to the next span of another three variables. This is what we call "simple projection". Each projection in the sequence is a 3D scatterplot of three variables. It is more than the 3D scatterplots, however, because more information could be unveiled by the animation moving from one projection to the next. Another straightforward way is random tour. By choosing randomly a 3D subspace and moving to the next randomly chosen 3D subspace, random tour creates a way for global dynamic browsing of multidimensional data. In the data preprocessing stage of a data mining project, simple projection and random tour are efficient ways to examine the distribution of values of each variable, the correlations among variables, and to decide which variables should be included in further analysis. Although real world databases have often many variables, these variables are often highly correlated, and databases are mercifully inherently low-dimensional. Simple projection and random tour are useful to identify the appropriate subspaces in which further mining is meaningful.

There are various ways of choosing interesting projections and projection sequences in a tour. For clustered data sets, one promising way is to use positions of data clusters to help choosing projections. Let us assume that a data set is available as data points in the p-dimensional Euclidean space and has been clustered into k clusters. Each cluster has a centroid which is simply an average of all the data points contained in the cluster. As we know, any four distinct and non-colinear points uniquely determine a 3D subspace. If we choose the centroids of any four clusters and project all data points onto a 3D subspace determined by these four cluster centroids, the Euclidean distance between any two of the four cluster centroids will be preserved and the four clusters will be displayed as separate as possible from each other. We call this a cluster-guided projection. Observe that there are $\binom{k}{4}$ possible cluster-guided projections. By using the grand tour to move from one cluster-guided projection to another, a viewer can have quickly a good sense of the positions of all data clusters.

There were both linear and nonlinear techniques[2] for dimension reduction of high dimensional data. Rather than nonlinear techniques such as Sammon's projection[15] which aims at preserving all inter-cluster distances by minimizing a cost function, we found linear projections more intuitive for the purpose of unveiling cluster structure and suitable for human eye's exploration. Linear projections and scatterplots could be found in many visualization systems (for example, the earlier Biplot[11]). The idea of using grand tour of lower dimensional projections to simulate higher dimensional displays was first proposed in [1]. Techniques were developed to design the path of a tour, for example, to principal component and canonical variate subspace[13], or to hill-climbing paths that follows gradients of projection pursuit indices[5,10]. An example visualization system which implements 2D projections and grand tour is XGobi[16]. For the visualization of data clusters, a 2D cluster-guided tour was proposed in [8].

To exploit human eyes' 3D nature of visual perception, we developed a visualization system for 3D projection and cluster-guided tour. A CAVE immersive virtual environment[6,7] is at our disposal for 3D immersive display. With the CAVE as a 3D "magic canvas", scatterplots can be drawn in mid-air in the 3D virtual space. This helps greatly data analysts visualize data and mining results. It helps to show 3D distributions of data points, locate similarity or dissimilarity between various clusters, and furthermore, determine which clusters to merge or to split further. Compared with other systems mentioned above, the grand tour in the CAVE virtual environment has characteristics such as: (1) 3D projection; (2) immersive virtual reality display; (3) cluster-guided projection determined by 4 data clusters; and (4) vary intuitive add-on tools for interaction and drill-down. It represents a novel tool to visualize multidimensional data and is now routinely employed for preprocessing data and analyzing mining results. It is also used to visually communicate mining results to clients.

The paper is organized as follows: Section 2 is to introduce grand tour . Section 3 discusses in detail the 3D cluster-guided projections and cluster-guided tour. Section 4 is for projection rendering inside the CAVE virtual environment. Section 5 presents add-on features such as projecting variable vectors together with data points, interactive picking and drill down, and cluster similarity graphs. Section 6 concludes the paper with future work and directions.

2 Grand Tour

For easy illustration, suppose we are to make a 2D tour in 3D Euclidean space (Fig. 1). A 2D oriented projection plan, or a 2-frame (a 2-frame is an orthonormal pair of vectors), can be identified by a unit index vector that is perpendicular to the plan. The most straight way to move from one 2D projection to the next is a sequence of interpolated projections to move the index vector to the next index vector on the unit sphere along a geodesic path.

For 3D grand tour of p-dimensional ($p > 3$) data sets, in the same way, it is necessary to have an explicitly computable sequence of interpolated 3-frames in p-dimensional Euclidean space. The p-dimensional data is then projected, in turn, onto the 3D subspace spanned by each 3-frame. For the shortest path to move from one 3D projection to another, the sequence of the interpolated 3-frames should be as straight as possible. Here "straight" means: If we think of the interpolated 3D subspaces as being evenly-spaced points on a curve in the space of 3D subspaces through the origin in Euclidean p-space (a so-called

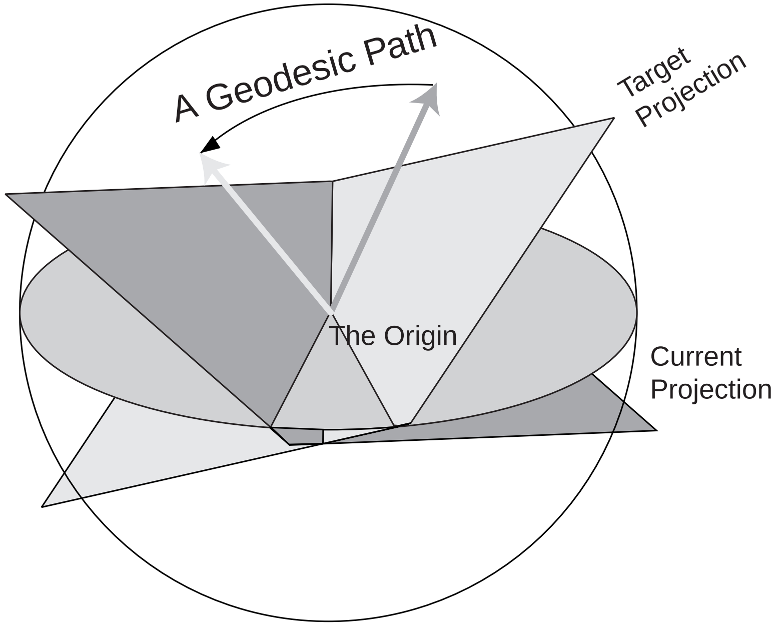

Fig. 1. Moving 2D projections along a geodesic path in 3D space.

“Grassmannian manifold”)(Fig. 2), we should be able to choose that curve so that it is almost a geodesic.

Moving along a geodesic path creates a sequence of intermediate projections moving smoothly from the current to the target projection. This is a way of assuring that the sequence of projections is both comprehensible, and also that it moves rapidly to the target projection. For 3D projections, a geodesic path is simply a rotation in the (at most) 6-dimensional subspace containing both the current and the target 3D spaces. This implies that some pre-projection is necessary in implementation so that computing data projections is within the joint span of the current and the next 3D subspaces, the dimension of which can be substantially smaller than p. Various smoothness properties of such geodesic paths are explored in great detail in [3]. For a description of implementation details, see [13, Subsection 2.2.1].

3 3D Cluster-Guided Projection and Cluster-Guided Tour

Let $\{X_i\}_{i=1}^n$ denote a data set, that is, a set of n data points each taking values in the p-dimensional Euclidean space R^p, $p > 3$. Let $X \cdot Y$ denote the dot product of two points X and Y. Write the Euclidean norm of X as $\|X\| = \sqrt{X \cdot X}$, and the Euclidean distance between X and Y as $d(X, Y) = \|X - Y\|$. Let us suppose

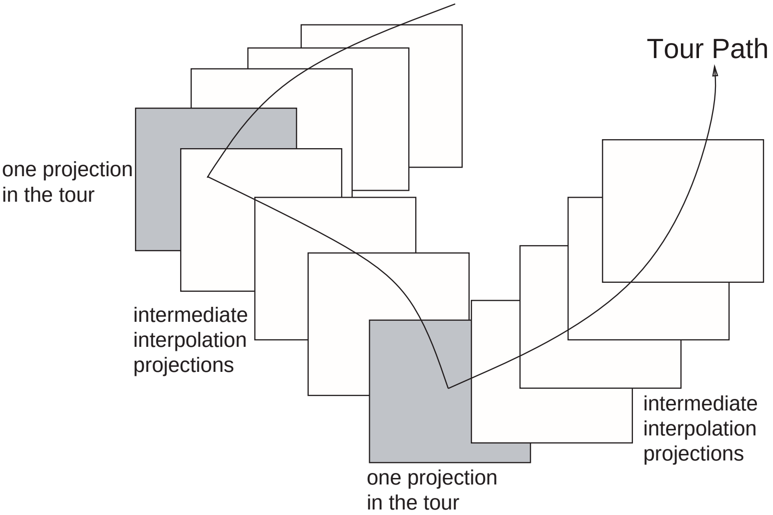
Tour Path
one projection in the tour
intermediate interpolation projections
intermediate interpolation projections
one projection in the tour

Fig. 2. A path of intermediate plans that interpolates a sequence of projection plans.

that we have partitioned the data set into k clusters, $k \geq 4$, and let $\{C_j\}_{j=1}^k$ denote the cluster centroids.

Any four distinct and non-colinear cluster centroids C_a, C_b, C_c and C_d in $\{C_j\}_{j=1}^k$ determine an unique 3D subspace in R^p. Let K_1, K_2 and K_3 constitute an orthonormal basis of the subspace (this could be obtained by orthonormalizing $C_b - C_a$, $C_c - C_a$, and $C_d - C_a$). We can then compute a 3D projection by projecting the data set $\{X_i\}_{i=1}^n$ onto the 3-frame (K_1, K_2, K_3). This projection preserves the inter-cluster distances, that is, the Euclidean distance between any two of the four cluster centroids $\{C_a, C_b, C_c, C_d\}$ is preserved after the projection. Specifically, let $X|p = (X \cdot K_1, X \cdot K_2, X \cdot K_3)$ denote the 3D projection of a p-dimensional point X, then $d(X|p, Y|p) = d(X, Y)$ for any $X, Y \in \{C_a, C_b, C_c, C_d\}$. This inter-cluster-distance-preserving projection is a right perspective of view that these four clusters are visualized as far as possible (Fig. 4).

There are various ways to choose the path (sequence of projections) of tour. One way is to simply choose a tripod from the variable unit vectors of p-dimension as the axes of one 3D projection and move from this projection to the next whose axes are another tripod. This is what we call "simple projection"(Fig. 3). It gives a way to continuously check a sequence of scatterplots of data against any three variables. Another straightforward way is random tour where each projection in the sequence is randomly generated. This gives a way for global dynamic browsing of multidimensional data.

Cluster-guided tour is a way to get cluster centroids involved in choosing projection sequences: Given k cluster centroids, there are at most $\binom{k}{4}$ combi-

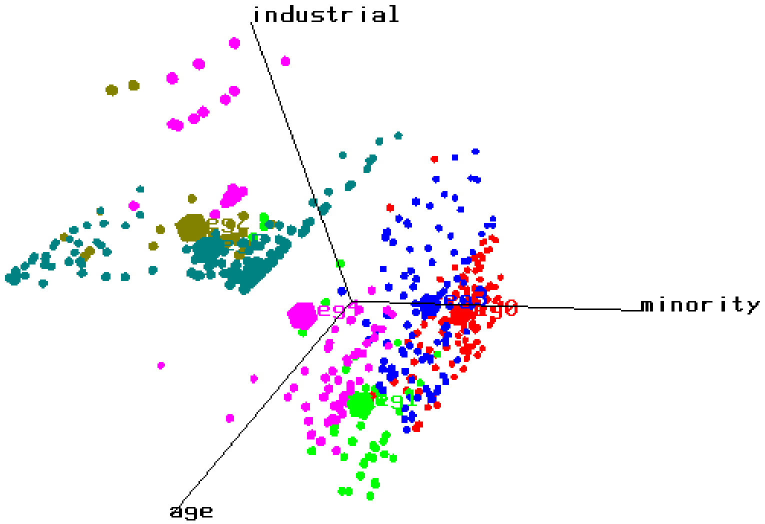

Fig. 3. Simple projection: a 3D scatterplot.

nations of unique 3D cluster projections. Each projection allows us to visualize the multidimensional data in relation to four cluster centroids. To visualize the multidimensional data in relation to all cluster centroids, we display a sequence of cluster-guided projections and use grand tour to move continuously from one projection in the sequence to the next.

The basic idea behind cluster-guided tour is simple: Choose a target projection from $\binom{k}{4}$ possible cluster-guided projections, move smoothly from the current projection to the target projection, and continue. We illustrate the 3D cluster-guided projection and guided-tour on the Boston housing data set from UCI ML Repository[4]. This data set has $n = 506$ data points and $p = 13$ real-valued attributes. The data set is typical (not in size, but in spirit) of the data sets routinely encountered in market segmentation. The 13 attributes measure various characteristics such as the crime rate, the proportion of old units, property tax rate, pupil-teacher ratio in schools, etc., that affect housing prices. We normalized all the 13 attributes to take values in the interval $[0, 1]$. To enable the cluster-guided tour, any clustering algorithm could be used to cluster the data set. Here we clustered the data set into 6 clusters by the Kohonen's Self-Organizing Map[14]. The six result clusters have 114, 46, 29, 107, 78, and 132 data points respectively. There are $\binom{6}{4} = 15$ possible 3D cluster-guided projections. We plot one of them in Fig. 4. To underscore the 3D cluster-guided

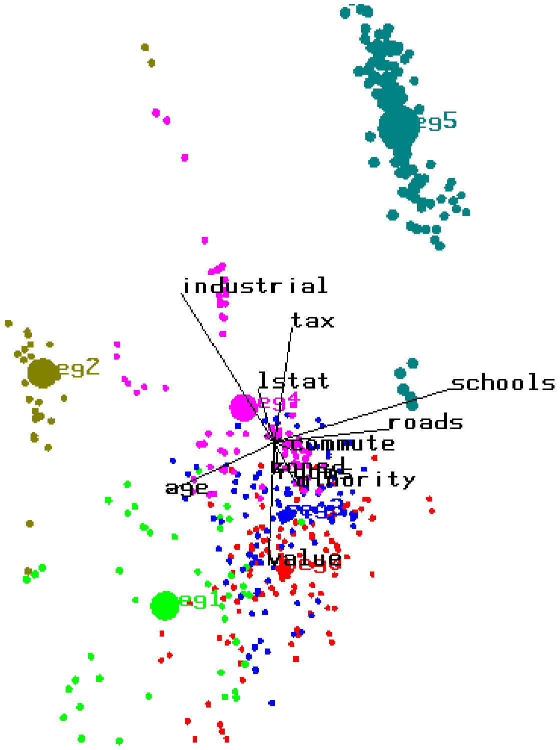

Fig. 4. A 3D cluster-guided projection determined by centroids (big balls with labels) of Clusters 1, 2, 4, 5. The four clusters are visualized as separate as possible. A p-pod of variable vectors is shown. Each ray of the p-pod represents the projection of a variable axis whose length represents the maximum value of the variable.

projections in locating interesting projections, compare Fig. 4 to Fig. 3 where we display a scatterplot of one of the attributes "industrial — proportions of non-retail business acres" against two of the other attributes "minority" and "ages of units." Unlike the scatterplot, the 3D cluster-guided projections reveal significant information about the positions of the clusters.

4 Rendering inside the CAVE Virtual Environment

CAVE is a projection-based virtual reality environment which uses 3D computer graphics and position tracking to immerse users inside a 3D space. The CAVE in IHPC has a 10 × 10 × 10 feet room-like physical space. Stereographic images are rear projected onto three side walls and front projected onto the floor. The four projected images are driven by 2 InfiniteReality graphics pipelines inside an SGI Onyx2 computer. The illusion of 3D is created through the use of LCD shutter glasses which are synchronized to the computer display through infrared emitters alternating the left and the right eye viewpoints. The CAVE allows multiple viewers to enter the CAVE and share the same virtual experience. But only one viewer can have the position/orientation of his/her head and hand captured.

With the CAVE as a 3D "magic canvas", 3D projection of high dimensional data is rendered as a galaxy in mid-air in the virtual space(Figure 5). The projection can be reshaped, moved back and forth, and rotated by using a wand (a 3D mouse). Each data point is painted as a sphere with its color representing the cluster it belongs. Spheres can be resized, and the speed of motion can be manually controlled anytime during a tour by adjusting an X-Y sensor attached on the wand. For easy identification, cluster centroids are painted as big cubs and labeled with cluster names. The variable vectors, which show the contribution to the projection of each variable, are visualized as lines in white color from the origin and marked by the names of variables at their far ends. There are two different ways of interactive picking: brushing with a resizable sphere brush; and cluster-picking by selecting a cluster's centroid. The CAVE has plenty of space for data rendering. At some future time, we may have multiple viewing projections synchronized and displayed simultaneously.

5 Add-On Features

5.1 Where We Are in a Tour?

A dizzy feeling besets many first-time viewers of high-dimensional data projections and they may ask "How do I know what I am looking at". In geometric terms, the task is to locate the position of a projection 3-frame in p-space. A visual way of conveying this information is to project the variable unit vectors in p-space like regular data, and render the result together with data points.

Examples of the application are shown through the Figures 3–5. A generalized tripod called "p-pod" is an enhanced rendition of the p variable unit vectors in p-space. Variable vectors in the p-pod can be treated as if they were real data, rendered as lines, and labeled by variable names in the far ends so that they are recognized as guide posts rather than data. In the figures, we choose the maximum value rather than the unit value of a variable as the length of its variable vector. The p-pod looks like a star with p unequal rays in 3D space, each indicating the contribution of a variable to the current projection.

5.2 Interactive Picking and Drill-Down

An advantage of grand tour is that an viewer can easily keep track of the movement of a certain group of data points during the whole journey of a tour. A cluster, or a set of data points, could be picked up by pointing to the cluster centroid or using a brushing tool. Data points picked up so far can be related back to the data, thus makes it possible for further analysis such as launching another mining process for drill down.

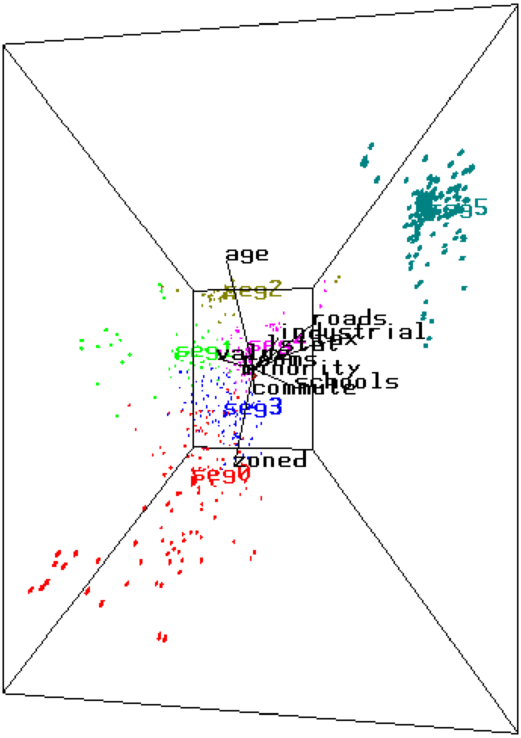

Fig. 5. This is a cluster-guided projection in the CAVE. The wireframe box indicates a 3D room where data points are plotted in mid-air. The cluster-guided projection is determined by centroids of Clusters 1, 2, 4, and 5.

5.3 Cluster Similarity Graphs

3D cluster-guided projection is continuous transformation of data. Two points which are close in R^p will remain close after projection. However, two points which are close in a 3D projection need not be close in R^p. There is a loss of information in projecting high-dimensional data to low-dimensions. To somewhat mitigate this information loss, we use cluster similarity graphs[9] as an enhancement to cluster-guided projection.

A cluster similarity graph can be defined as follows. Let vertices be a set of cluster centroids $\{C_j\}_{j=1}^{k}$, and add an edge between two vertices C_i and C_j if $d(C_i, C_j) \leq t$, where t is a user-controlled threshold. If t is very large, all cluster centroids will be connected. If t is very small, no cluster centroids will be connected. It is thus intuitively clear that changing the threshold value will reveal distances among cluster centroids. The cluster similarity graph can be overlaid onto the projections. For example, straight lines connecting the cluster centroids in the Fig. 6 represent a cluster similarity graph at a certain threshold. It can

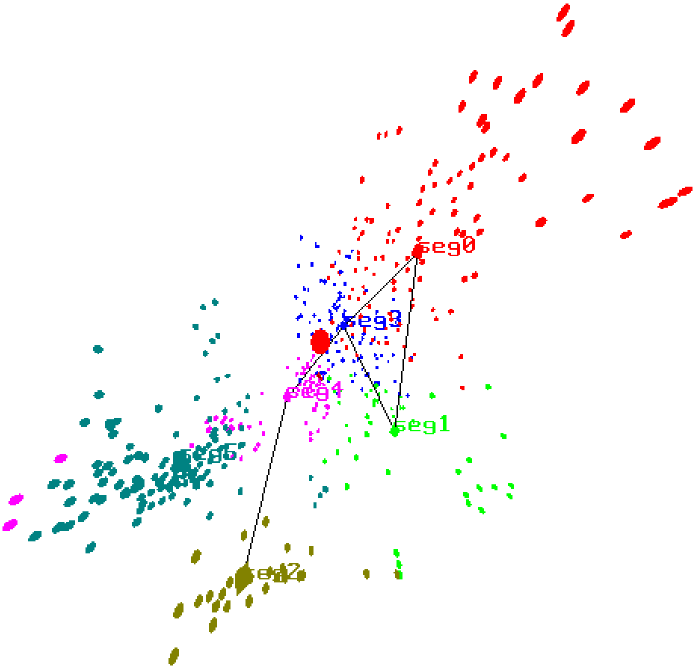

Fig. 6. An similarity graph adds yet another information dimension to cluster-guided projections.

be seen that the Clusters 0, 1, 3 are close to each other, among which Cluster 3 is close to Cluster 4 which is close to Cluster 2. Cluster 5 is a standalone cluster from all others. The cluster similarity graph adds yet another information dimension to cluster-guided projections, and hence, enhances the viewing experience.

6 Conclusion and Future Work

This paper discussed the use of 3D projections and grand tour to visualize higher dimensional data sets. This creates an illusion of smooth motion through a multidimensional space. The 3D cluster-guided tour is proposed to visualize data clusters. Cluster-guided tour preserves distances between cluster centroids. This allows us to fully capture the inter-cluster structure of complex multidimensional data. The use of the CAVE immersive virtual environment maximizes the chance of finding interesting patterns. Add-on features and interaction tools invite viewer's interaction with data.

The cluster-guided tour is a way to use data mining as a driver for visualization: Clustering identifies homogenous sub-populations of data, and the sub-populations are used to help design the path of tour. This method can also be applied to the results generated by other data mining techniques, for instance, to identify the significant rules produced by tree classification and rule induction. All these are possible ways to allow a user to better understand both results of mining and data at hand.

One important thing about an algorithm is its scalability. Grand tour scales well to large data sets. Its computational complexity is linear to the number of variables. The number of variables matters only in calculating projections, i.e. dot products, which has a linear complexity to the dimensionality of arguments. There are two major steps in grand tour, calculating a tour path and making projections. Calculating a tour path is nothing with the total number of data points. Making projections has a computational complexity linear to the number of data points. This is in the sense that all data points have to be projected one by one. For large data sets, this complexity can be greatly reduced by making density map instead of drawing points.

The following directions is being explored or will be explored in the future:

- *Working with categorical variables.* In relational databases it is quite common for many of the variables to be categorical rather than numerical. A categorical variable can be mapped onto a linear scatterplot axis in the same way as a numeric variable, provided that some order of distinct values of that variable is given along the categorical axis. Categorical values may be explicitly listed. The order of the values being listed will be the order these values be arranged on the axis. Categorical values could be grouped together, reflecting the natural taxonomy of values. Categorical values could also be sorted alphabetically, numerically by weight, or numerically by aggregate value of some other variable. We are working on having categorical variables involved in a tour, and some results may come up soon.
- *3D density projection and volume rendering.* Scatterplot loses its effectiveness as the number of points becomes very large. It has also a drawback that identical data records may coincide with each other. For a tradeoff between computational complexity, comprehensibility and accuracy, we plan to use dynamic projections of high dimensional density map as a model to visualize data sets which contain large number of data points. 3D density projection is important to study, especially when clusters are not balanced in size and when clusters overlap with each other. Research is now on finding solutions of problems such as: how to store the sparse, voxelized high dimensional data more efficiently; and how to fast render a volume of high dimensional voxels onto the projected 3-dimensional space.
- *Parallel implementation for better performance.* A parallel implementation is necessary for the rendering of very large data sets. Since data points are independently projected, it should be quite straightforward to parallelize the code, for instance, by using multithreads on a shared memory machine. Since our CAVE's backend computer, the SGI Onyx2, is quite busy with CAVE display, leaving few resources for projection calculation, a client-server implementation is also necessary. This will be done through a high speed

network connection to a more powerful SGI Origin2000. All projection data will be calculated on the server and sent in real time to the CAVE. One interesting issue here is how to transfer only the necessary projected data to the CAVE in order that the transferred data can be directly rendered.

References

1. D. Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM Journal of Science and Statistical Computing*, 6(1):128–143, January 1985.
2. G. Biswas, A. K. Jain, and R.C. Dubes. An evaluation of projection algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-3:702–708, 1981.
3. A. Buja, D. Cook, D. Asimov, and C. Hurley. Theory and computational methods for dynamic projections in high-dimensional data visualization. Technical report, AT&T, 1996. http://www.research.att.com/~andreas/papers/dynamic-projections.ps.gz.
4. E. Keogh C. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.
5. D. R. Cook, A. Buja, J. Cabrera, and H. Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 2(3):225–250, 1995.
6. C. Cruz-Neira. *Projection-based Virtual Reality: The CAVE and its Applications to Computational Science*. PhD thesis, University of Illinois at Chicago, 1995.
7. C. Cruz-Neira, D. J. Sandin, T. DeFanti, R. Kenyon, and Hart J, C. The cave audio visual experience automatic virtual environment. *Communications of the ACM*, 35(1):64–72, 1992.
8. I. S. Dhillon, D. S. Modha, and W. S. Spangler. Visualizing class structure of multidimensional data. In *Proceedings of the 30th Symposium on the Interface: Computing Science and Statistics*, Minneapolis, MN, May 1998.
9. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.
10. J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23:881–890, 1974.
11. K. R. Gabriel. Biplot display of multivariate matrices for inspection of data and diagnois. In V. Barnett, editor, *Intrepreting multivariate data*, pages 147–173. John Wiley & Sons, New York, 1981.
12. P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–474, 1985.
13. C. Hurley and A. Buja. Analyzing high-dimensional data with motion graphics. *SIAM Journal on Scientific and Statistical Computing*, 11(6):1193–1211, 1990.
14. T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, New York, second extended edition, 1997.
15. J. W. Sammon, Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18:401–409, 1969.
16. D. F. Swayne, D. Cook, and A. Buja. XGobi: Interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998.