

Data visualization

Zhao Kaidi

School of Computing,

National University of Singapore

iscp0075@nus.edu.sg zhaokaidi@hotmail.com

Matrix Number: HT00-6177E

(Document Version 1.0)

Abstract

Data visualization is a quite new and promising field in computer science. It uses computer graphic effects to reveal the patterns, trends, relationships out of datasets. In this paper, we first get familiar with data visualization and its related concepts, then we will look through some general algorithms to do the data visualization. To get deeper about it, we will have some discussion about multidimensional data visualization. With the combination of some known methods, we present a new algorithm to do 4 dimensional data visualization. We also present a program project plan about it (optional), and some issues and explanations around it.

Introduction

Human has a long history with basic data visualization, and data visualization is still a hot topic today. The history of visualization was shaped to some extent by available technology and by the pressing needs of the time, they include: primitive paintings on clays, maps on walls, photographs, table of numbers (with rows and columns concepts), these are all some kind of data visualization – although we may not call them under this name at that time.

Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. It is also the process of transforming objects, concepts, and numbers into a form that is visible to the human eyes. When we say “information”, we may refer to data, processes, relations, or concepts. Here, we restrict it to data.

Data visualization is all about understanding ratios and relationships among numbers. Not about understanding individual numbers, but about understanding the patterns, trends, and relationships that exist in groups of numbers [4].

From the point of user understanding, it may involve detection, measurement, and comparison, and is enhanced via interactive techniques and providing the information from multiple views and with multiple techniques.

Why do we do data visualization?

To see and understand pictures is one of the natural instincts of human, and to understand numerical data is a years training skill from schools, and even so, a lot of people are still not good with numerical data [4]. From a well-drawn picture, one is much easier to find the trends and relations. Because visual presentation of information takes advantage of the vast, and often underutilized, capacity of the human eye to detect information from pictures and illustrations. Data visualization shifts the load from numerical reasoning to visual reasoning. Getting information from pictures is far more time-saving than looking through text and numbers – that's why many decision makers would rather have information presented to them in graphical form, as opposed to a written or textual form.

Another thing we should mention is that: data visualization is NOT scientific visualization. Scientific visualization uses animation, simulation, and sophisticated computer graphics to create visual models of structures and processes that cannot otherwise be seen, or seen in sufficient detail. While data visualization is a way that presents and displays information in a way that encourages appropriate interpretation, selection, and association. It utilizes human skills for pattern recognition and trend analysis, and exploits the ability of people to extract a great deal of information in a short period of time from visuals presented in a standardized format.

Background

Before we focus on multi-dimensional data visualization, let's review some basic concepts of data visualization and graphical technology.

Talking about graphics, we should remind what is called graphical entities and attributes. When visualizing, generally we only have the following graphical entities and attributes to select from (although not limited to):

Entity: point, line(curve), polyline, glyph, surface, solid, image, text

Attribute: color/intensity, location, style, size, relative position/motion

What we call “data”, actually have some special characters, and often can be divided into following groups [6]. They include (but not limited to):

Numeric, symbolic (or mix): 123, or @

Scalar, vector, or complex structure:

Various units: meters, inch.

Discrete or continuous: 1, 2, 3, or p

Spatial, quantity, category, temporal, relational, structural

Accurate or approximate

Dense or sparse

Ordered or non-ordered

Disjoint or overlapping

Binary, enumerated, multilevel

Independent or dependent

Multidimensional, etc.

We consider the data is properly visualized, if the visualization is [6]:

Effective: viewers can interpret it easily.

Accurate: sufficient for correct quantitative evaluation.

Efficient: minimize data-ink ratio and chart-junk, show data, maximize data-ink ratio, reduce non-data-ink, reduce redundant data-ink

Aesthetics: must not offend viewer's senses

Adaptable: can adjust to serve multiple needs

Data Visualization Techniques

Bearing the above in mind, we have some commonly used representation ways in data visualization, they include (but not limited to):

Charts: bar or pie

Graphs: good for structure, relationships

Plots: 1- to n-dimensional

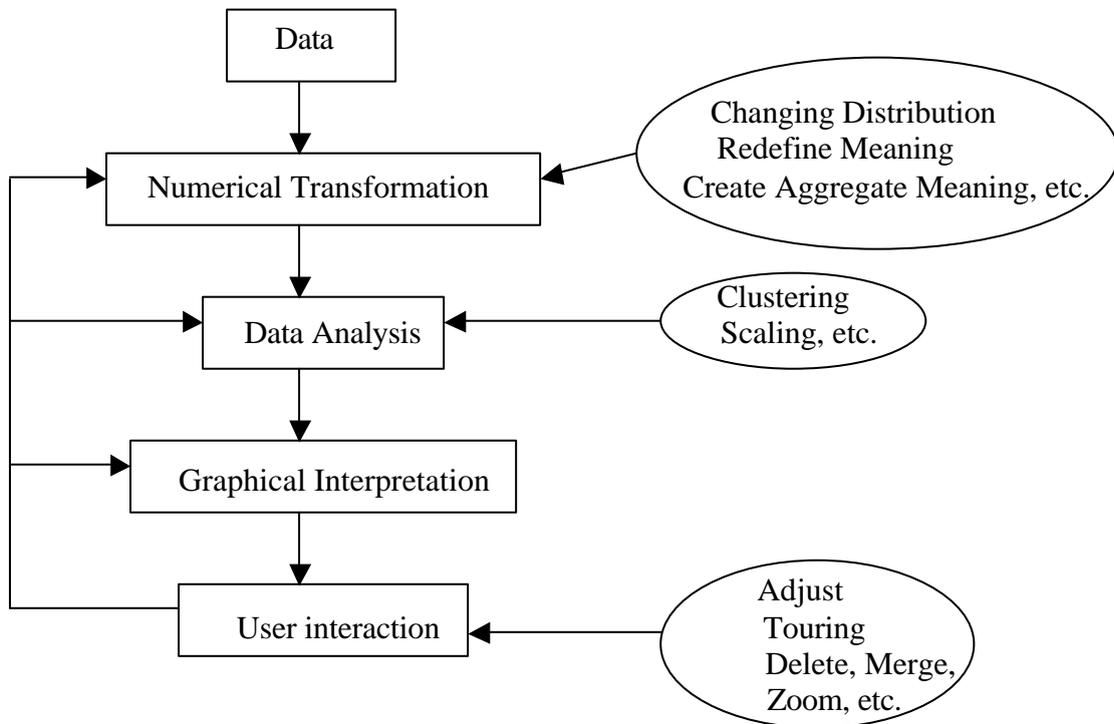
Maps: one of most effective

Images: use color/intensity instead of distance (surfaces)

3-D surfaces and solids

Isosurfaces/slices

We also have some common steps in data visualization [4], they include:



Numerical Transformation:

Visualization is a kind of transformation of numerical data. Numbers are abstract concepts, and to represent them as points and lines requires a transformation.

Transformations include:

1) Changing the distribution: modify the distribution of numbers so that they are more suitable for analysis or visual presentation. Some frequently used ways include:

Linear transformation

Logarithmic transformation

Normalizing transformation

Arcsin transformation

Square root transformation

Inverse transformation

2) Redefining the Meaning: adjust numbers so that they are more meaningful, or more representative of the concept that the data analyst is interested in.

3) Creating (new) Aggregate Meaning

Data Analysis

Data analysis is the process of applying various methods to data to assist in interpretation. Some of the exploratory data analysis methods are: statistical support, cluster analysis, multidimensional scaling, and factor analysis. Data analysis can be used to transform data or to summarize the data itself or its statistical properties.

Graphical Interpretation

Graphical interpretation consists of a few key activities such as judgment of magnitude (and relative magnitude), judgment of proportion (and relative proportion), judgment of trend and slope, and judgment of grouping. It may also use some ways such as:

- Use scaling and offset to fit in range

- Use derived values (residuals, logs) to emphasize changes

- Use projections, other combinations, to compress information, get statistics

- Use random jiggling to separate overlaps

- Use multiple views to handle hidden relations, high dimensions

- Use effective grids, keys and labels to aid understanding

User Interaction

When presented with the visualization results, users may find it does not fit their minds properly. Users may want to do the followings, which may require to re-do some earlier steps.

- Dynamically adjust mapping

- Tour data by varying views

- Labeling to get original data

- Deleting to eliminate clutter

- Brushing/Highlighting to see correspondence in multiple views

- Zooming to focus attention

- Panning to explore neighborhoods

Multi-dimensional Data Visualization ($N \geq 3$)

Most of the data visualization methods are taken from the old days when paper publish industry dominates the world. As they originate from paper publish, they take use of the papers as a media, which is a 2-dimension media. They handle 2-dimension

data quite well. They also can present basic 3-dimension data with the help of projection. But when it comes to high dimensional data, these data visualization methods no longer stand.

When talking about N-dimensional ($N \geq 3$) data visualization, we have several ways to do it. They are:

Translate to N-1 dimensional data for visualization

Use special viewing instrument, such as stereoscope

Use special viewing methods, such as stereograph

Use ordinary 2/3-dimensional algorithms plus various attributes to represent data in other dimensions.

Use animation.

We will exam these ways one by one. We will mainly focus our attention on 4-dimensional data visualization if applicable.

Translate to n-1 dimensional data

This is the most used ways to handle $N \geq 3$ data visualization. The basic method for it is projection. Given a N-D data, we may first project it into (N-1)-D, if required, we also can continue our projection into (N-2)-D, until we can properly handle the visualization.

For example, if we are to visualize data set of (x, y, z), we use (x, y, z) as three axis to get a 3-D object, given the sumption that we have some way to visualize this 3-D object for our data examination (parallel viewing projection or perspective viewing projection, for example).

Another method of translate N dimensional data to N-1 is also invested. [1] In case where one dimension has a very limited range it is possible to effectively combine two dimensions, A and B into a single dimension C. We can call this multiplexed dimensions.

Use special viewing instrument

One kind of special glasses are called “stereoscope”. With this glasses, when looking 2 purposely prepared color pictures, one can see a 3D image out of these two 2D pictures.

With the help of this, we can visualize our 3 dimensional data into 3D objects, and present them on two 2D pictures, while viewers can get the 3D image easily.

Use special viewing methods

As I shall claim that it is human that is the user (viewer) of the result of data visualization. So, scientists are trying to find whether human eyes, without any other assistant, can directly get some N-D information out of some (N-1)-D data representation. The research is successful in 3D, but (I shall say) “very limited”. By purposely adjust our eyes’ focusing point, we can get a 3D image out from a 2D paper image without any other aid. This is sometimes called “stereograph”.

But the big drawbacks are: it is not a convenient way. And while some people can do this, some people just can’t.

Use ordinary 2/3-dimensional algorithms plus various attributes to represent data in other dimension

Simply put, $4=3+1$ ($N= m + i$, $m<N$). The basic idea is to present a 3-dimensional representation showing some of the points of the object, and to add the fourth dimension information to it. There are two broad ways to do it, one is using some graphical attributes to present our fourth dimension data, the other one is using some non-graphical attributes to do so.

Using graphical related attributes to present the fourth dimension:

This is a very natural way, because when we perceive information, we are so used to use our eyes. Basically, we have several methods to do so.

Color

Color is an important and frequently-used feature in data visualization. In multidimensional data visualization, for example in 4D, we can use the first 3 dimensions to construct a set of 3D objects, then we use different color for the points to indicate the difference in the fourth dimension data. One thing to point out it, the data point can either be in a 2 dimensional plane, or be a 3 or higher dimensional space. When doing so, questions arise: How many colors can we use in one view, and how can we choose effective colors that provide good differentiation between data elements during the visualization task? From our experience, we know that if too many colors are used, the view become awful for us to get any useful information – we just could not easily figure out the data points among different colors. As for the question of choosing effective colors, we can get a simple method from [8].

Size of the point

Much like the color above, size as an attribute of data point is also important in data visualization. One thing to mention: the “point” here is actually an “object”, not the same as “point” in mathematics. Same as above, the data point can either be in a 2

dimensional plane, or be a 3 or higher dimensional space. Several main drawback of size include: it may be difficult to implement in conjunction with any projection that introduces a decrease in size with apparent distance. And, the size of each data point will greatly reduce the total number of data points in one view.

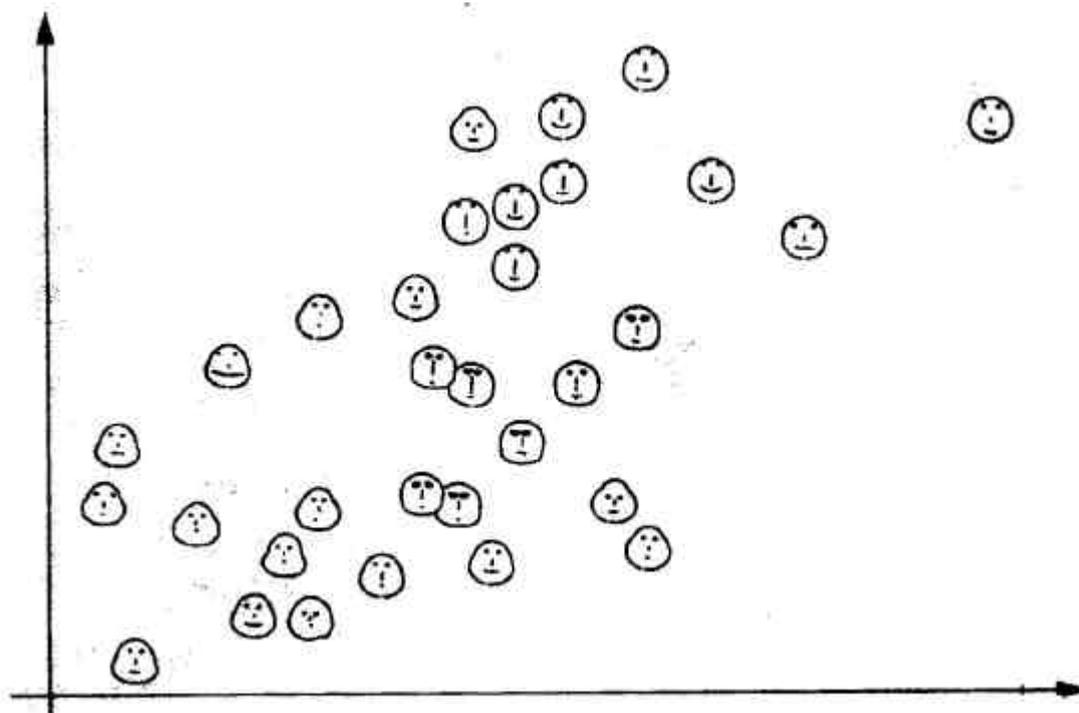
Glyphs (Iconic) Displays

Glyphs (also referred to as icons) are graphical entities which convey one or more data values via attributes such as shape, size, color, and position. [9] [14] [19]The use of glyph attributes for data visualization is based on human perceptual abilities.

Based on what attribute is used, there are three most commonly used glyphs.

Chernoff Faces

This is a widely-accepted way of visualize multidimensional data. A chernoff face is a special type of glyph that capitalizes on human sensitivity to faces and facial features. [4] [3] See the figure below (this is a case with view of 2D plane, if in 3D space, it also stands, or we can use the other methods as described later.).

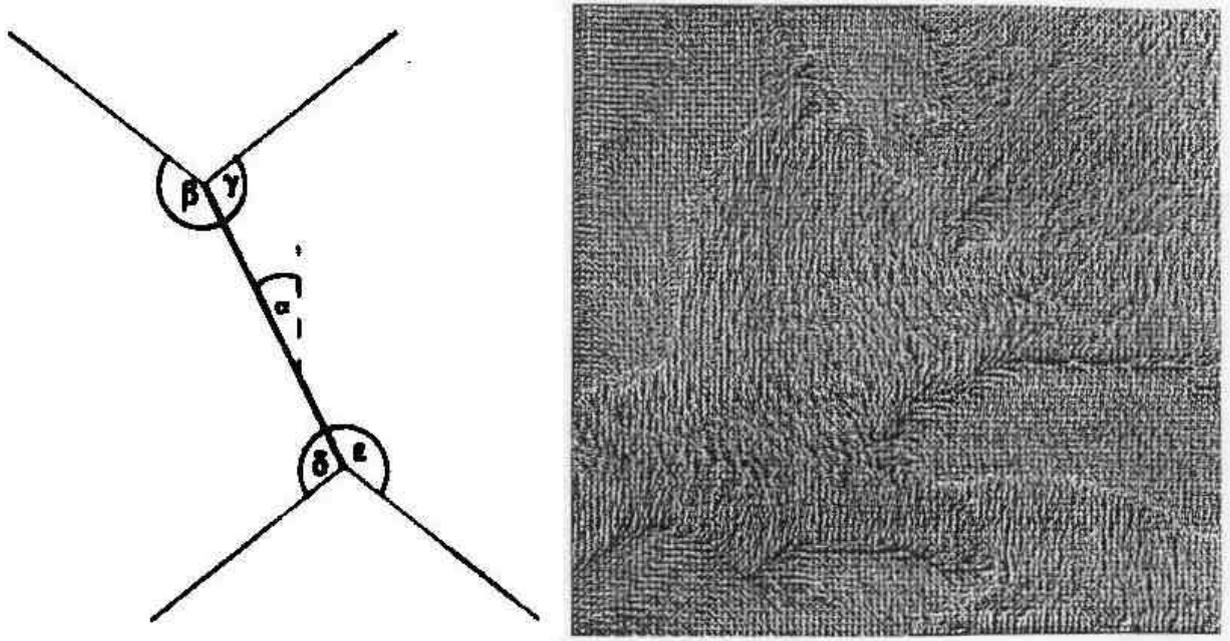


In this case, there is one face that stands out from the other faces. As this example shows, “data” faces can be useful in detecting outliers (strangers) and in grouping data visually.

Stick Figure

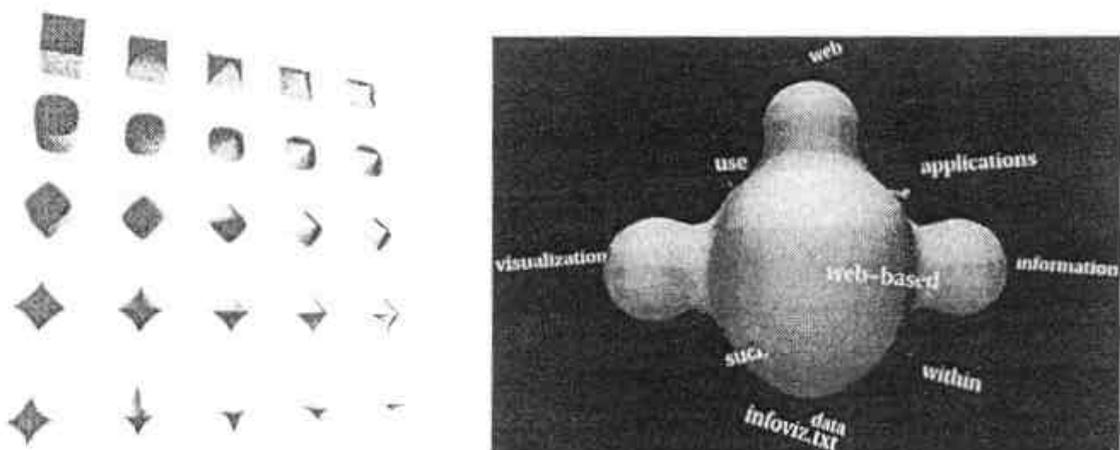
Another well-known way is called the “stick figure technique” [3]. Here, the icon is some type of stick figure, where angle and /or length of the limbs represent the data dimensions. The figure below is an example of stick to visualize 7 dimensions.

When the data items are relatively dense with respect to the display dimensions, this results a texture patterns that vary according to the characteristics of the data. See the figure below. Generally, the individual icons are not intended to be discernible.



Shape (Glyph) Displays

In 3D space, using the shape of a data point (object) is a more challenging but more effective data visualization way, because three-dimensional shape perception is not as well understood as color, size, and spatiality perception. Another problem is the design of meaningful glyphs. Below is a result from the study of [2].



(Left: Example superquadric shapes.)

(Right: Example implicit surface shape for a document on “web-based information visualization”)

Vector originating from point

Another well studied way is to use a vector originating from the data point to represent an additional data dimension. The direction and / or the length of the vector are used to represent this additional information. [20] shows a small example where 4D data has been presented in 2D: The single points in the 2D space have each been given two "legs". The size of the legs represents the value of the point in the last two dimensions.

Using non-graphical related attributes to present the fourth dimension

Sound

As the popularity of computers and computer multimedia, another way is recently under research: use sound to present the higher dimension data. This is mentioned in [10] [18]. Some principles are: the sound should not interfere with the visualization process; it should be simple as most of us can understand with the combination of the visual information, but not too simple that becomes unsatisfactory and annoying; should be some interactive, users can have some control.

Use animation

Actually, a lot of high dimensional visualization (especially 4D) programs (especially in engineering) use animation to achieve their purpose. In animation, with the moving (changing) 3D visualization, the fourth dimension – time, is also represented. Examples include [13] [15] [16] [17].

One of my opinion: if none of the dimensions is time, this method can also be adopted by carefully selecting one attribute / dimension and transform it as the “time” in the dataset (how are other dimensions change according to this attribute?).

My 4-D Data Visualization Algorithm: Dynamic Color Plane

My specified 4-dimensional data visualization algorithm (called “dynamic color plane” multidimensional data visualization algorithm) takes the advantage of human eyes’ complex functions toward colors. We use a color-based visualization algorithm. And we also take the advantage of animation in our algorithm.

We use a dynamic-color-plane to move through the 3D data points space, and dynamically change the colors of the points that interact with the plane. We also can change the moving direction of this plane, allowing the user to study any relations between the points and the colors. By doing so, the user may effectively figure out the distribution of the colors, and the tendency of the data, which may be very helpful to the viewer.

This algorithm is superior to the existing color-based algorithms, for the reasons:

In some conditions as will be mentioned in next section, my algorithm acts as the general visualization algorithms that take use of colors. So my program would be at least as good as these visualization programs.

In other conditions when the dense data points prevent visualization effects in the generally algorithms, my algorithm will be better because we take use of: 1) clustering technology. 2) our specified dynamic color plane algorithm moving.

By the way, any general algorithms that assist visualization can also be intergraded into my algorithm.

My Project Program (Optional)

(This program is optional to the course project)

My proposed program based on my algorithm to visualize 4-dimensional data should be:

Given 4D datasets as (X, Y, Z, W) , my implication is constructing a set of 3D point (dots) set in the 3D space, and using another way (color) to represent the fourth attribute. My proposed steps include:

1) Search in X, Y, Z, W respectively, find the one that most suitable to act as the fourth dimension that we can use color to represent. (For example it is attribute W .)

When selecting, we may consider these ways:

A) We try to select an attribute that has fewest possible values. If so, we can represent it using different colors directly.

B) If all of the attributes have many possible values, then we select an attribute that can be clustered into fewest numbers of clusters. Then we can use colors to represent the clusters.

C) If all the attribute is linear, we then just select one attribute and bin it. Or, if one attribute has the similar characters as “time”, we can consider it as “time” and use animation based on it.

2) Construct the 3D space using (X, Y, Z).

Before this step, we may consider doing some numerical translation, as mentioned before. About the user interface, user may want to be able to rotate this 3D world by mouse.

3) If the total number of data points is small, we then can give colors to the data points directly. But if there are too many data points, then we shouldn't color them, or else, they overlap and we can't get any idea of the data visualization. If so, we paint them as semi-transparent, and we will use a dynamic way.

4) Select a “start point” and an “end point” to form a dynamic color view direction.

How to select these two points is a question. Mouse can only move in the 2D screen, and we have seen many applications try to “guess” the user's intended mouse position in the 3D space using some algorithms. We can use the mouse button to let the user indicate which direction he wants in the 3D space. For example, if no button is pressed when moving, the movement is in the X-Y plane. If the right button is pressed when move, the user indicates this moving is along the Z direction.

5) Construct a semi-transparent plane that moves according to the color view direction. Color the dots in this plane as the plane interacts them. The color is calculated from the (clustered) W attribute. Un-color the dots when the plane has passed.

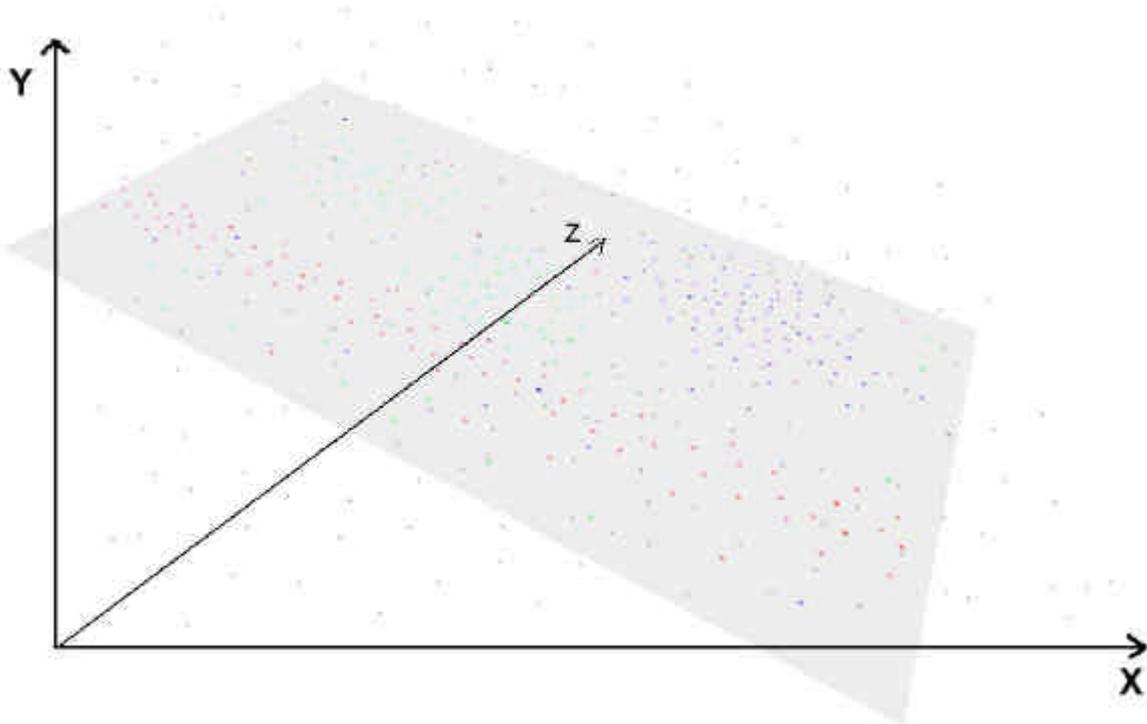
6) User can select different “start point” and “end point” so that user can see different relations that X, Y, Z have with attribute W.

7) If more than one attribute is suitable to act as W, then we should allow the user to visualize them by color as the W attribute.

8) Visualization result.

As our color plane moves through the data points, we may find color planes or color line or color clouds, each may be a interest to the user. We can help the user to find them by counting the distance between the dots, and display it at user's requests.

This following is an illustration of this program project.



We have an alternate way about the color. If step one, if we find out that the attribute only has relatively limited possible values, or can be clustered into relatively limited possible values, then we can also adopt the shaped glyph method to represent them. This leads our representation similar to [2].

Discussion

This paper presents a wide knowledge on the field of data visualization. Actually, from my research and survey, I find that the main point in data visualization is not merely on “how to” visualize the data, but also on how can make the visualization useful and understandable to viewers. From the relatively easy 2D visualization to deeper multi-dimensional visualization, the most important thing we consider is how to make the visualization meaningful to users, and how the viewers can get what they want from this visualization. From this point, the data visualization is actually helping the viewer to mining the data.

Due to limited resource, my proposed visualization program (optional to the course research) is not complete yet. But from the theoretical analyze of my visualization algorithm, my visualization program should be superior to the general ones, for the reasons that when it suits, it acts like the general color-based algorithms, otherwise, it

uses clustering and/or a dynamic representation. It also can adopt any generally algorithms to assist the visualization.

Reference

- [1] Shaun Bangay, “Visview: A system for the visualization of Multi-dimensional data”, in “Visual Data Exploration and Analysis V”. (TA 1505 Pse 3298)
- [2] David S. Ebert, Randall M. Rohrer, Christopher D. Shaw, Pradyut Panda, James M. Kukla, D. Aaron Roberts, “Procedural Shape Generation for Multi-dimensional Data Visualization”, in “Data Visualization ‘99”. (QA 90 Jei)
- [3] Daniel Keim, “Visual Support for Query Specification and Data Mining”. (QA. 76.9 Dbm. kd)
- [4] Kamran Parsaye, Mark Chignell, “Intelligent Database Tools & Applications”. (QA 76.9 Dbm.PS)
- [5] Barbara Hausmann, Geometry Center, “Viewing Four-dimensional Objects In Three Dimensions”, from www.geom.umn.edu/docs/forum/polytope/
- [6] Matthew Ward, “Overview of Data Visualization”, from www.cs.wpi.edu
- [7] Steven Richard Hollasch, “Four-Space Visualization of 4D Objects”, from <http://www.research.microsoft.com/~hollasch/thesis/default.htm>
- [8] Christopher G. Healey, “Choosing Effective Colours for Data Visualization”, from www.cs.ubc.ca/libs/imager/tr/healey.1996b.html
- [9] Finn Årup Nielsen , “Visualization and Analysis of 3D Functional Brain Images”, from <http://hendrix.imm.dtu.dk/staff/students/fnielsen/thesis/finn/finn.html>
- [10] Dennis Roseman, “What Should a Surface in 4-Space Look Like?”, in “Visualization and Mathematics”. (QA 90. Vis)
- [11] Alfred Inselberg, “Multidimensional Detective”, in “IEEE Symposium on Information Visualization 1997”. (TK 7882 Inf. Iv3)
- [12] Pak Chung Wong, R.Daniel Bergeron, Gergory M. Nielson, Hans Hagen, Heinrich Muller, “Scientific Visualization, over view, methodologies, techniques”. (Q175 Nie.)
- [13] Laboratory for Scientific Visual Analysis, <http://www.sv.vt.edu/index.html>, <http://www.sv.vt.edu/classes/ESM4714/exercises/exer9/exer9.html>

- [14] Matthew O. Ward, “A Taxonomy of Glyph Placement Strategies for Multidimensional Data Visualization”, from <http://davis.wpi.edu/~matt/courses/glyphs/>
- [15] Helena Mitsova, Lubos Mitas, Bill Brown, Irina Kosinovsky, Dave Gerdes, Terry Baker, John Isaacson, “Interpolation and Visualization from 3D and 4D scattered Data Using GRASS GIS” from <http://www2.gis.uiuc.edu:2280/modviz/viz/vol1.html>
- [16] Rich Balfour, application of “fourDviz”, from <http://www.infinity-technologies.com/itweb/html/4d.html>
- [17] some 4D examples, from <http://www.stanford.edu/group/4D/sections/4d-examples.htm>
- [18] U. Axen, I. Choi, “Using Additive Sound Synthesis to Analyze Simplicial Complexes”, in “Proceedings of 1994 International Conference on Auditory Display”.
- [19] Jim Foley and Bill Ribarsky. “Next-generation Data Visualization Tools”. In L. Rosenblum, R. A. Earnshaw, J. Encarnacao, H. Hagen, A. Kaufman, S. Klimenko, G. Nielson, F. Post, and D. Thalmann, editors, “Scientific Visualization, Advances and Challenges”. (T385 Sci)
- [20] K. V. Mardia, J. T. Kent, and J. M. Bibby. “Multivariate Analysis”. (QA278 Mar)