VisuMap — A High Dimensional Data Visualizer

Version 3.0

Introduction

Most industries nowadays maintain operational and historical data with relational database in form of tables. Data mining relational databases has become increasingly critical for enterprises in order to exploit new opportunities and respond to new challenges. Decision makers have often to study those data before a mature data model is available. In those cases, visualization is often the most effective investigation method.

VisuMap is a unique software application specially developed to help people to explore fresh data to discover new knowledge. VisuMap unleashes the perception power of experienced human eyes, and enables people to quickly analyze tables with large number of rows and columns from different perspectives.

The Challenge

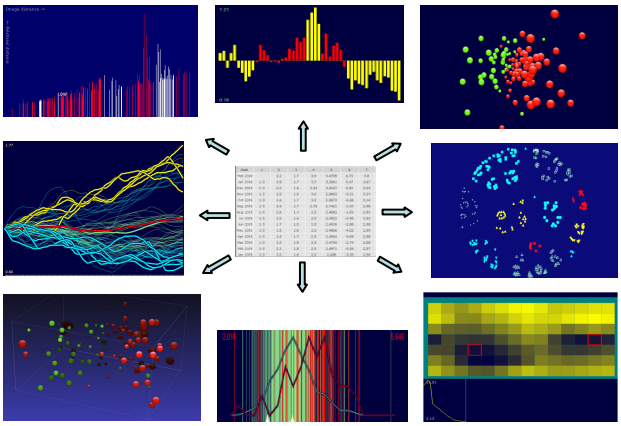
A major challenge for visualization method is the high dimensional nature of many data tables: If a table contains more than 3 data columns a visualization method has to convert the high dimensional data to 2 dimensional graphics with certain means.

Most available visualization software relies on the user to select few data columns or calculate few attributes for the visualization purpose. Those software applications share the common limitation: they all require good knowledge about the data.

Some software try to alleviate the limitation with interactive exploration support; and some others resort to more generic multidimensional scaling methods. Yet, there is no software on the market that treats the *dimensionality reduction* as a central task. Thus, exploratory data analysis of true high dimensional data remained so fare an ad-hoc and difficult undertaking.

The Solution

With a *visual centric design* VisuMap is a software application developed from ground up to target the analysis of high dimensional data. After data has been imported into the application, data points will be represented by dots, glyphs, curves or bars in various *visual data views*. Each of those data view offers a different perspective of the data. Those views are linked with each others; so that a user can simultaneously investigate a dataset from different perspectives with multiple views.



At its core VisuMap implements a collection of *dimensionality reduction* algorithms which map high dimensional data to 2D or 3D maps. Tables with large number of columns of numerical and textual data can be quickly transformed into maps which give the user an overview of the whole dataset.

On top of those core services VisuMap offers whole palette of features to support interactive visual data analysis. Those features include, among the others, data clustering, data editing and annotation, data drilling, data sorting, intelligent data import/export and fast 3D animation.

For advanced users VisuMap offers a script interface to automate frequently performed

tasks, as well as a library plug-in interface to implement domain specific extensions.

Applications

VisuMap has found successful commercial applications in the following areas:

- Pharmaceutics industry
- Telecom industry
- Financial data analysis
- Marketing data analysis
- Network traffic analysis
- Bioinformatics and Healthcare Service.

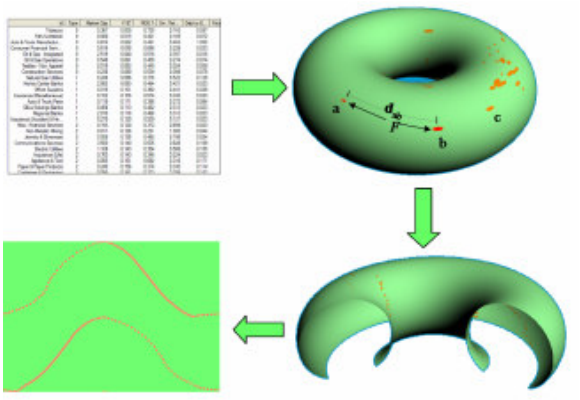
Main Services of VisuMap

Dimensionality Reduction

A table in a relational database can be considered as a high dimensional dataset by viewing each data record as a data point in a high dimensional space whose dimension is the number of columns of the table. One of the core services of VisuMap is to *map* the high dimensional datasets to 2 or 3 dimensional spaces while preserving relevant information.

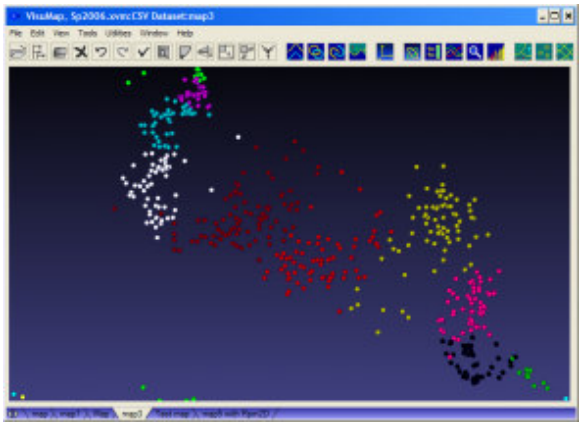
Since different applications require different information set, it is critical to provide different methods to preserve different type of information. VisuMap implements the following dimensionality reductions algorithms:

- **Principal Components Analysis** (PCA). A fast linear projection algorithm. This method offers a fast high level overview of a dataset; it is good for simple mainly linear datasets. But this algorithm may be inappropriate for data with non-linear properties.
- **Sammon Mapping**: A widely used non-linear mapping algorithm. This method offers a fairly good overview of the global characteristics of the dataset while provide only limited capability to preserve non-linear features.
- **Curvilinear Component Analysis** (CCA). A new non-linear mapping algorithm with time dependent strategy. This method preserves local non-linear features much better than the previous methods. This method is relatively more calculation intensive.
- **Relational Perspective Map** (RPM). The latest non-linear mapping algorithm that is based on simulation of dynamic system on closed manifolds. This method is able to perform global segmentation of the dataset and present local features in a non-overlapping manner. This method is relative calculation intensive.



RPM Algorithm

VisuMap implements all algorithms in a consistent way, so that the user can easily generate different maps from the same dataset and quickly compare them from different perspectives. VisuMap user interface is designed like a modern spreadsheet application so that it is easy to use to for people with experiences with spreadsheet applications. The following picture shows a snapshot of VisuMap user interface:



On an up-to-date personal computer, VisuMap allows user to create a map for up to 8000 data points. For fast 3D animation VisuMap employed the latest hardware level 3D library (Managed DirectX).

Metric and Filter

VisuMap implements the concept *metric* and *filter* to offer maximal flexibility in defining information set on datasets. Whereas the dimensionality algorithms focus on mapping high dimensional data to low dimensional space, the metric and filter mechanisms enables the user to directly specify different information set on the same dataset. For instance, if a user is more interested in the correlation between data points, he/she can use the correlation metric instead of the normal Euclidean distance.

VisuMap currently supports the following metric:

- Euclidean distance.
- Mahalanobis distance.
- Symmetrical information.
- Hamming distance.
- Direct distance matrix.
- Pearson correlation
- Speaman ranking correlation
- Kendall ranking correlation.
- Graph tree distance.
- Intersection distance
- Wave hedge distance
- Arbitrary custom metric implemented through the VisuMap plug-in interface.

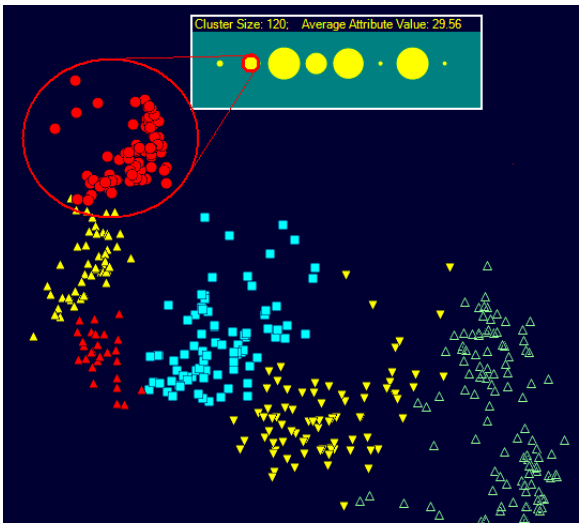
The filter concept allows users to quickly enable, disable or scale data columns. Filters in VisuMap are implemented as stand-alone objects, so that users can share and copy filters between datasets.

Data Clustering

VisuMap implements a collection of clustering algorithms with which the user can generate clusters and annotate their representation with different colors. The clustering algorithm also allows user to reduce dataset size by generate new data points from clusters. VisuMap is able to cluster dataset with more than 500 000 data points, and reduce datasets to appropriate size for interactive investigation.

Currently VisuMap supports the following clustering algorithms:

- **K-Mean Clustering**: a classic clustering algorithm for multidimensional dataset.
- **Agglomerative Clustering**: an algorithm to cluster datasets with any well defined distance matrix.
- **Self-Organizing Map** (SOM): a clustering algorithm that preserves topological properties underlying a dataset. The self-organizing map is especially appropriate to reduce size of large dataset.
- **Self-Organizing Graph** (SOG): a proprietary extension of the SOM algorithm that performs data clustering according to arbitrarily structured network.
- **Metric Sampling**: a special clustering algorithm for non-Euclidian datasets, e.g. datasets in form of tree or network structures.
- **Affinity Propagation**: a recent clustering algorithm for generic data equipped with similarity matrix.

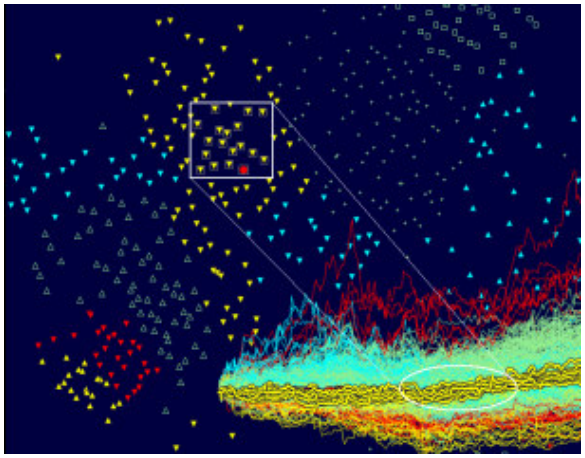


Clustering and Coloring dataset
with K-Mean Clustering Algorithm

Interactively Linked Views

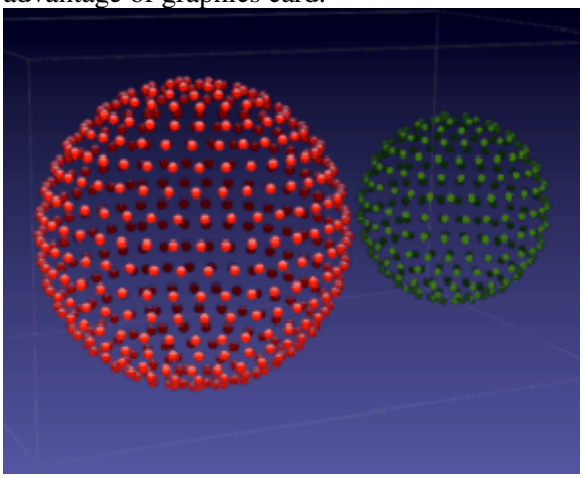
With VisuMap users can explore a dataset from different perspectives simultaneously with multiple views. The data points are represented as glyphs, curves or bars in different views. Changes made in one views will be immediately reflected in other views.

The following picture shows a dataset with two views — the RPM map view in which each data point is represented as a colored dot; and the value curve view in which each data point is represented as colored curve. When the user select a subset of data point in the RPM map, their corresponding curves will immediately highlighted in the curve view.



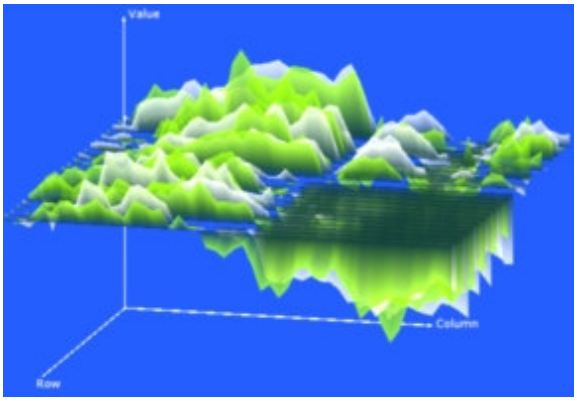
VisuMap implements the following views:

- The scatter plot view in which a data point is represented as a glyph. The distance between the glyphs will be generated by one of the dimensionality reduction algorithms.
- The curve plot view in which each data point is represented as curve.
- The details view that represents the whole dataset as a table with one row for each data point showing the details of the data point. Special user interface optimization has been implemented so that the user can directly load tables with up to 10 000 columns into the application.
- The table view that presents the complete dataset as an editable table.
- Zoom view which allows user to zoom-in into sub-regions of a scatter plot view.
- Shepard diagram that plots the relational distance in high dimensional space against the image distance in 2D/3D space.
- 3D animation view that enables users to navigate in a 3D space. The 3D animation view provides fast animation by taking advantage of graphics card.



3D animation view

- Attribute map that shows the data attributes (e.g. data columns) as scatter plots. The user can then study the relationship between the attributes and perform attribute related operations.
- Spectrum view that displays data of a selected dimension as spectrum style map.
- Bar chart view that displays data of a selected dimension as a bar chart map.
- Multibar chart view that displays multiple series of data in a bar chart window.
- Mountain view that displays a complete data table in a mountain landscape 3D style.



Mountain View

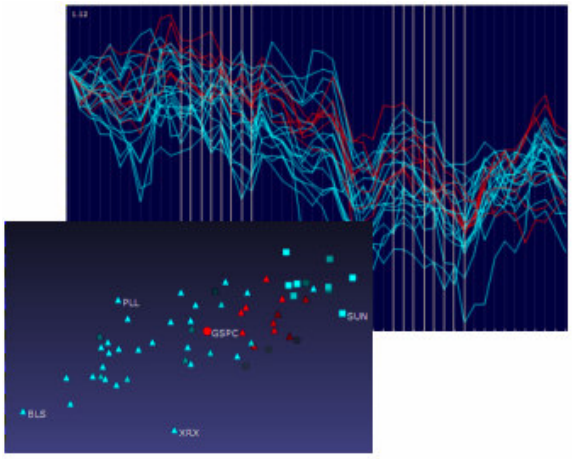
- Atlas view that enables user to organizer and compose visual information of different types in a single window.

Dual Mode Data Drilling

For the investigation of complex large dataset it is prevalent to be able to focus the investigation on subsets of data. VisuMap provides more than a dozen of methods to enable users to select data interactively and visually. Users can then successively drill into data subsets and explore them with any of the available views.

Moreover, most views in VisuMap can operate in either data point mode or attribute mode. In the data point mode the user drills into the dataset by selecting data points which correspond to rows in the dataset table. In the attribute mode the user drills into the dataset by selecting attributes which correspond to columns in the dataset table.

The attribute mode provides a quick way to explore patterns and regularities underlying subset of attributes. For example, the following curve plot view shows the daily price history of some traded stocks. The diagram indicates two time intervals in which the stocks experienced collectively down-turns. If we are interested in correlation between these stocks during these two periods we can simply select those attributes in attribute mode and open the PCA window to study them in 3D view.



Programming Interfaces

VisuMap provides two types of programming interfaces for advanced usages:

- The *script interface* enables users to automate most interactive tasks with script written in standard JavaScript language. VisuMap offers simple GUI user interface to create, test, execute and integrate scripts. The script interface can also be used to interact with third party applications. For example, a user can create a script to send selected data to Microsoft Excel to use its data rendering services.
- The *plug-in interface* enables advanced users to extend VisuMap with domain specific features. Plug-in modules are libraries implement in any language supported by the .Net platform. For instance, for analyzing graphic data we can implement the Laplacian kernel through the plug-in interface. The plug-in interface also offers full access to the script interface.

For some major industries, like the financial industry, plug-in modules are available from VisuMap Technologies.

Data Import and Export

Substantial effort has been put into VisuMap to enable quick and easy interaction with third party applications. With few mouse clicks, for instance, a user can copy and paste data from/to applications like Microsoft Excel or image editors.

VisuMap currently support the following data importing methods:

- Importing data text files or clipboard in comma separate vector (CSV) format. VisuMap is able to detect a large number of variations of the CSV format.
- Importing data from relational database through the ODBC interface.
- Importing data from tables embedded in web pages. The user can mostly import data directly from a web page through the copy-and-paste operation.

- Importing images through copy-and-paste operation. Most popular image formats (like JPEG, GIF, etc.) are supported.

VisuMap current offers following data exporting methods:

- Exporting data to clear text file in CSV format. A user can copy any subset of a dataset to clipboard through a single Ctrl-C key pressing.
- Exporting data to XML format controlled by XSLT script.
- Exporting data and map to SVG (scalar vector graphic) format.
- Exporting image and diagrams to files or the clipboard. Most popular image formats (like JPEG, GIF, etc.) are supported.
- VisuMap uses XML as its dataset format. Third party applications can access native VisuMap dataset directly according the well designed XML-Schema.

An Example

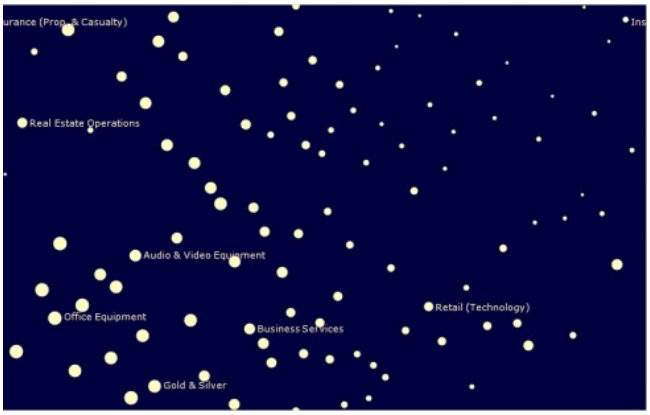
A typical use scenario of VisuMap starts with a tabular dataset. As an example, the following picture shows the key financial ratios of 104 US industry sections in Nov. 2003. Each industry is characterized by 8 numerical values, e.g. market capitalization, dept ratio, etc.

Id:Industry	Market Cap	P/E	ROE %	Div. Yield %	Debt to E...	Pric
Advertising	55.47	26.721	12.37	1.023	1.63	
Aerospace & Defense	143.02	29.033	5.578	1.901	1.118	
Air Courier	23.81	29.243	11.129	0.261	0.29	
Airline	51.97	32.24	4.014	0.12	2.01	
Apparel/Accessories	48.05	20.558	21.584	1.273	0.334	
Appliance & Tool	25.91	18.163	34.01	2.466	2.053	
Audio & Video Equipment	125.81	35.698	-10.266	0.16	0.527	
Auto & Truck Manufacturers	353.84	12.187	14.388	3.482	11.861	
Auto & Truck Parts	53.44	16.321	11.927	2.135	0.784	
Beverages (Alcoholic)	138.9	21.157	57.153	1.699	2.332	
Beverages (Non-Alcoholic)	242.26	24.056	34.109	1.584	0.549	
Biotechnology & Drugs	512.44	41.199	4.097	0.363	0.292	
Broadcasting & Cable TV	339.31	40.233	-0.673	0.958	0.759	
Business Services	182.53	31.309	16.467	1.096	0.422	
Casinos & Gaming	37.56	25.153	18.409	1.757	1.794	
Chemical Manufacturing	156.44	25.063	12.917	2.075	0.834	
Chemicals - Plastics & Rubber	106.26	55.973	3.293	3.423	1.375	
Coal	11.2	27.964	5.544	2.573	1.155	
Communications Equipment	506.1	42.4	5.263	0.838	0.239	
Communications Services	1517.53	17.662	19.98	4.778	1.321	
Computer Hardware	281.53	28.804	28.128	0.712	0.52	

VisuMap can help us, for instance, quickly answer these questions:

- Which industry sections are similar to each other?
- Is an investment portfolio sufficiently diversified?
- How to maximize the future return of an investment portfolio?

With VisuMap we can quickly create a single map for the 8 dimensional dataset as be shown in the following picture:



In above map each spot represents an industry section. The size of spots reflects the average price/earning ratio of the industry sections: larger spots represent industry sections with large price/earning ratio. A main property of this map is that two industries with similar financial ratios will be mapped to closely located positions. Thus, an investment portfolio is well diversified if its corresponding industry sections are widely distributed across the whole map.

From the map we can quickly notice that closely located industry have similar spot size. This means two industries with similar financial ratios will have similar P/E ratio. Any exceptions to this relationship should get special attentions.

System Requirements

- Windows 2000/XP/Vista/Winodws 7
- Microsoft .Net 2.0 Runtime Library.
- Microsoft DirectX Runtime Library Dec. 2005 Release (required only for 3D animation).
- DirectX 9 compatible graphic card (required only for 3D animation).
- 2GHz CPU with 500 MB or more main memory.

Conclusion

VisuMap offers a unique and powerful collection of services for visual exploratory analysis of high dimensional dataset. By focusing on the goal of dimensionality reduction and by employing the latest software technologies VisuMap achieved a flexible, high performance and user friendly tool for data mining and knowledge discovery.

Contacts

James X. Li, Ph.D.,
CEO & Chief Scientist
252 Edgehill Dr. N.W.
Calgary, Alberta T3A2W8

Phone: +1 403 547-9630
Mobile: +1 403 607-8240
Email: information@visumap.net
http://www.visumap.net