Interactive Visualization of High Dimensional Marketing Data in the Financial Industry

Ruud Smeulders, Anton Heijs
Rabobank Group, Treparel
r.j.a.m.smeulders@rn.rabobank.nl , aheijs@treparel.nl

Abstract

We describe the visualization of high dimensional marketing data for a financial asset management company. The data typically consists of 30 to a 100 variables of 25000 to half a million clients. We use the visualization of the correlation matrix as a variable selection tool which makes it easier to find patterns in the data. The user can then select data ranges of the selected variables and start a cluster analysis using 5 variables.

The clustered data are then visualized as a set of spheres. In an additional visualization we first sort data values of a client variable and then visualize the sorted cubic data in a cube using volume rendering and isosurfaces.

The interactive correlation visualization allows marketing researchers to quickly explore all kinds of combinations of variables, which enables them to find valuable client behavior patterns much faster. The cluster visualization allowed researchers to identify detailed groups of customer with similar behavior. Additionally, the visualization of the sorted cubic data gives in dept information of one variable over the total sample of customers. With these visualizations, a better understanding is given on customer behavior.

Keywords--- **Data mining, visualization**.

1. Introduction

The business decisions of today and more so of tomorrow depend on a better interpretation of data. To be in control companies collect large amounts of data and store these in databases and data warehouses. Information systems are used to compute even more data to gain a deeper insight in the original data. Obviously, the real value of the data is to obtain knowledge about the business.

One of the data intensive areas where knowledge and insight has to be absolutely up-to-date is marketing research. With a short time to market and a fast changing customer demand access to all the data in an intuitive way is vital. In database marketing studies large amounts of high dimensional data must be analyzed. Marketing databases often contain a large number of variables. It is very difficult to find combinations of variables which identify customer behavior patterns using only statistical methods. Without any a priori knowledge about the data records, several techniques are used to investigate the data. Traditional data-mining techniques are normally mostly related to statistical techniques, neural networks or genetic algorithms. All these techniques lack the possibility to provide some mechanism to explain their results.

Visualization methods are becoming increasingly important to analyze and explore these large multidimensional data sets [1-7]. This is sometimes also called visual data mining. MineSet [8-10] and Iris Explorer [11] are good examples of software tools for this approach. An important advantage of visualization methods over other data mining techniques is the direct interaction and immediate feedback for the user as well as visual steering directly on the original data [6, 8, and 13]. Here we present a case where visualization techniques are used to reveal the important variables in marketing datasets from an accounting system of a Dutch financial asset management company. We investigate and propose a 3D visualization approach to obtain insight by human pattern recognition in marketing data. In section 2 we describe the case and in section 3 our visualization approach, where after in section 4 we discuss the obtained results.

2. Case study

The database we used consisted of marketing data of 25000 customers (without name or address). The client data-set consists of 100 variables for each of the 25000 clients. There are 2 types of variables: data from the large customer database and additional data derived from a questionnaire. In the questionnaire 8 questions were send out to the 25000 clients of a total of 500000 clients. The data was analyzed and a logit model was used to predict entrepreneurship of the clients based on their answers. The logit model could also handle incomplete data for those situations where clients did not respond to all 8 questions. Using the logit model we acquired a new

Proceedings of the Ninth International Conference on Information Visualisation (IV'05)
1550-6037/05 $20.00 © 2005 **IEEE**

variable, entrepreneurship, which was to be analyzed in relation with the variables already available in a database called Rogiro. The type of variables in the customer database can be segmented into several groups; questions, countries, miscellaneous and fund variables.

In our visualization experiments we included the new variable entrepreneurship from the logit model. We only visualized variables with strong relationship to each other and to the entrepreneurship to improve the logit model.

3.1 Interactive 3d correlation matrix for sorting and selecting

Selecting the right set of variables from a dataset containing N variables is a difficult task. When N is large, the number of combinations for 2 variables is at least ½ N(N-1) and for combinations of 3 variables this is ½(N-1)N^2. The linear correlation coefficient represents a relevant measure of relationship between two metric datasets [12]. The linear correlation coefficient has a value between -1 and 1, whereby 0 represents no correlation. We propose to visualize the correlation of N variables in the dataset by calculating all N^2correlations of this set and plot them as bars on a NxN grid (see also [12]). The datasets is not restricted to contain only scalar data. One can visualize in a similar matter vector or tensor datasets. We used IRIS-explorer to selectively read in the data, which are then converted into an Iris-explorer lattice [11].

By calculating the correlations between all variables and visualizing the correlation matrix as a 3D bar landscape the user immediately has an overview of the most relevant variable combinations.

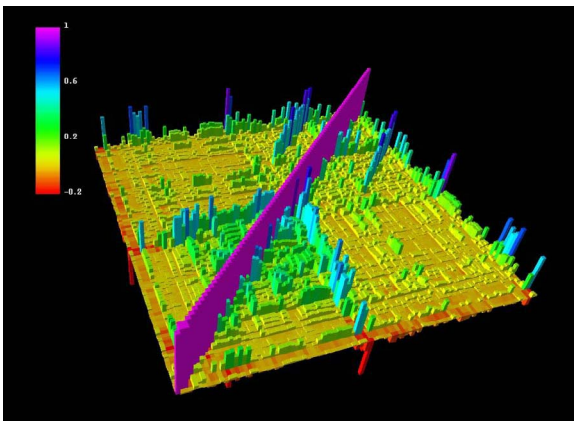

Figure 1 Correlations landscape of 100 variables for 25000 customers of marketing database.

Figure 1 shows the correlation landscape of the marketing dataset we used. In the landscape each bar represents the correlation of two variables for the total population of 25000 customers. Providing more textual information when a user points with the mouse over a correlation bar is essential, and some textual information and drill down facility are for this reason available.

The main goal of the correlation landscape is to provide insight in which variables have strong correlation to give marketing analysts the opportunity to select those variables in an interactive manner. By pointing on a 3D bar the user can visually select the two variables. The data distribution of these variables is then visualized in an adjacent window. Figure 2 shows an example of a three dimensional visualization of the client variables “age” and “relationship time”.

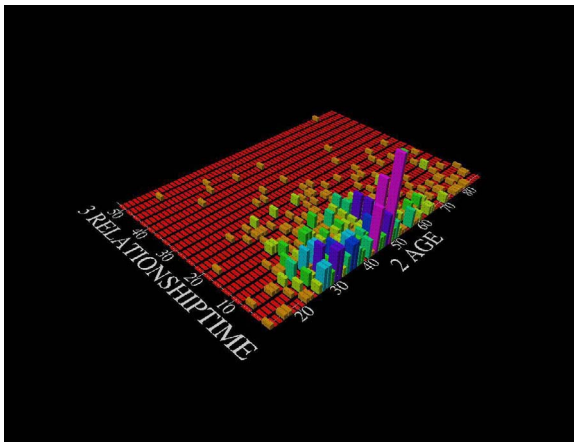

Figure 2 Visualization of data distribution of 2 variables which were selected using picking from the correlation matrix in figure 1.

3.2 Visualization of custom properties in spheres and colors

After the user has determined the data interval ranges from the data distribution visualization of the two variables, the next step is a cluster analysis. Apart from the two selected variables one can add three other variables for the clustering and analyze them in relation with the earlier selected variables. Similar customers will be clustered to form a sphere. The spheres are located in 3D space where the first two variables (from the correlation landscape) are on the x- and y-axes. The third variable will be the z-axis, while a fourth variable determines the size of the spheres and the fifth variable determines the color coding of the spheres. We used the number of clients belonging to a cluster to determine the size of the spheres.

Figure 3 gives an example of this visualization of the amount of customers with strong correlation for 3 variables (now x, y and z-axis). The spheres are ‘colored’ with the average value of the entrepreneurship of the clients in the clusters. The coupling between the multiple views, as described by Roberts et al [14], makes the exploration of patterns in the data more comprehensible.

Proceedings of the Ninth International Conference on Information Visualisation (IV’05)
1550-6037/05 $20.00 © 2005 **IEEE**

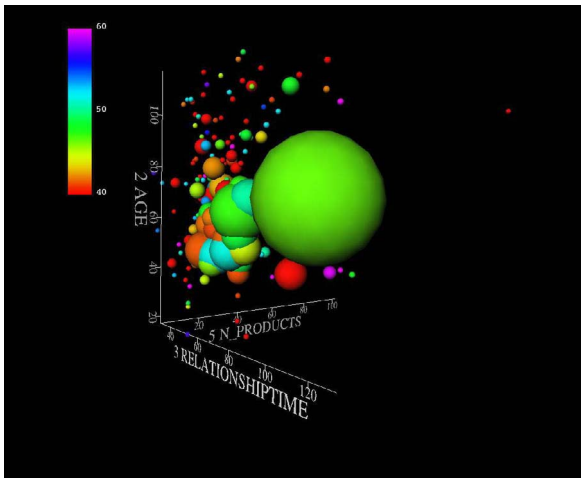

Figure 3 Similar customers are clustered into spheres; the radius of each sphere is determined by the number of customers with almost similar value of the 3 variables. The colors of the spheres represent the average value of the entrepreneurship. The variables "relationship time", "number of investment products" and "age" are on the x y and z axis.

3.3 Cube with sorted custom properties

The sorting is also essential in the way we visualize correlations between clustered data. If we have to deal with the variables "age" and "country", for instance, whereby we divided age in 10 groups and country in 44 different countries, sorting by country and the visualizing the correlation of classes of age for each country is totally different from sorting by age and then visualizing for each age group the variable country.

This was the start of a new way of visualizing in the 2nd data visualization. Now we started with sorting one of the variables found in the correlation visualization.

After the sorting we visualized the customers in a cube, starting at customer 1 with the lowest value of this variable in x, y, z equal to 0,0,0 down in the left bottom of the cube. Than, with the rising of the value for the particular variable, the second customer was put into 1,0,0. The 3rd into 2,0,0 and so on. The customer with the highest value can be found in the upper right corner of the cube. If the value of a second variable is almost similar for several customers, these customers are clustered into a sphere. A big sphere means that a lot of customers have about the same value for this (other) variable. The result can be found in Figure 4.

4 Discussion

One of the goals of the research was to explore client behavior in relation to entrepreneurship using the marketing database data and the logit model. After visualizing the 100 variables of the marketing database of 25000 customers in the correlation landscape, the

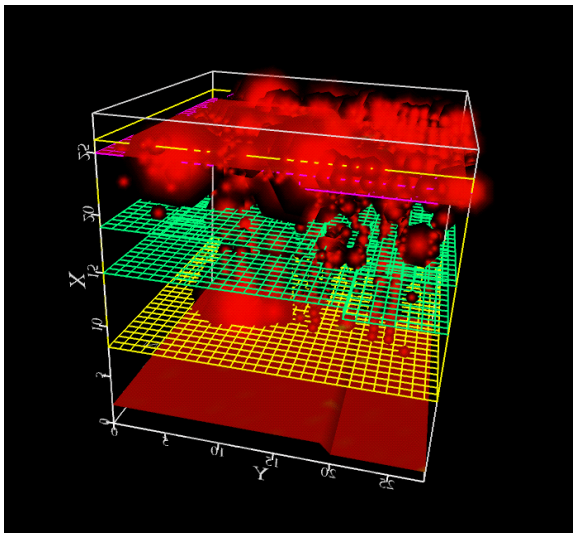

Figure 4 Clustering after sorting 25000 customers. Now, customers are sorted with 1 variable (age) starting at x, y, z = 0,0,0 and than 'growing' along x, y, z-axis. The clustering originates around amounts of customers with (almost) similar values for a second variable (asset).

marketing researchers easily picked out several combinations of variables to obtain the second visualizations. The correlation matrix was interactive in two different ways. By pointing on a bar the 2 variables were given with textual information. Here also some background information and drill down facilities were available. The marketing researchers appreciated this interaction and used the additional information to improve the choice of variable pairs. By picking a specific bar the marketing researchers selected the two variables for the second visualization. In figure 2 a simple visualization of the data distribution for the variables age and relationship time is shown as bars along the z-axis. Although this is a very simple form of visualization, marketers used this to improve there earlier decision to select interesting data ranges. Furthermore, they could improve there models for calculating "entrepreneurship".

In figure 3, clusters at positions in space according to related variables are visible. The variables in this visualization are: "relationship time" as x-axis; "number of investment products" as y-axis; "age" as z-axis. The two variables "age" and "relationship time" were selected because they showed a strong correlation in figure 1.

In the visualization of figure 3 we now see the clusters of customers in certain regions of the variables, where this correlation is very strong. Because now the real value of importance of each variable in combination with other variables is visible, marketers could further optimize their model for entrepreneurship. The colors of the spheres represent a certain value of one of the variables that was connected with the answers on the questionnaire. With this color visualization it was easier

to combine variables from the questionnaire to the model for entrepreneurship.

In figure 4 the customers were sorted by age. Thereafter clustering showed the amounts of customers with almost similar assets. Here, clusters of customers occur at certain levels of the variable age, where customers have certain range of assets. One should realize that in this visualization one can see every individual customer in the total dataset of 25000 customers. Every customer has an unique position in the cube of figure 4. This position is only depending on the value of the variable age. The spheres now indicate that several customers have almost the similar value for the variable asset.

Although interpretation of the visualizations is a hard job, the system with coupled visualizations is giving much additional information to marketers for the mutual relation of different variables. It generates new information that is not available without these visualizations. In this way a model is optimized to predict customer behavior and especially to predict entrepreneurship of customers from other variables of the customer database. The resulting model of the new variable entrepreneurship is not available for publication, due to its confidential character.

Conclusions

Visualization techniques are a good additional analysis tool for exploring high dimensional marketing data by marketing researchers. Other then often used techniques they give a better insight into the original data. By using the image pattern recognition possibilities of the human brains, visualization gives a powerful extra analyzing tool for data mining and abstract model making in marketing research.

Using the correlation landscape the marketing researchers could identify important correlation patterns in their data. By selecting variables with interesting correlations in an additional second visualization unknown relationships between clusters of variables are discovered. This resulted in new models to predict customer behavior.

Acknowledgements

This work was performed for the Marketing Research department of the Robeco Group in the Netherlands. The Robeco Group is part of the Rabobank Group. For more information please contact the corresponding author Anton Heijs (e-mail aheijs@treparel.nl at Treparel). The authors wish to acknowledge Gerard Wolfs and Erika Slagter from the Robeco Group.

References

[1] Ben Schneiderman, Tree visualization with treemaps: A 2 dimensional space filling approach, ACM, Transactions on Graphics, 11, pp 92-99, 1992.

[2] D.A. Keim, Visual Techniques for exploring databases. Invited tutorial Int. Conf. on Knowledge discovery in databases, Newport Beach, CA, 1997.

[3] Antony Unwin, Requirements for interactive graphics software for exploratory data analysis, Computational Statistics, 14, pp 7-22, 1999.

[4] Buja, A., Swayne, D. F., Littman, M., and Dean, N., XGvis: Interactive data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics.* 2001.

[5] D. Keim, Information Visualization and Visual Data Mining, IEEE Transactions on Visualization and Computer Graphics, 7(1), pp 1-8, 2002.

[6] Benjamin B. Bederson, Ben Shneiderman, Craft of Information Visualization, The: Readings and Reflections, Morgan Kaufman, Paperback, 2003.

[7] Antony Unwin and Graham Wills, Exploring Time Series Graphically, Artifical Computing and Graphics Newsletter, 2 pp 13-15, 2003.

[8] SGI Inc., MineSet (tm): a system for high-end data mining and visualization., Int. Conf. on very large data bases (VLDB'96) , Bombay, India, pp 595, 1996.

[9] B.G. Becker, Visualizing decision table classifiers, in proceedings IEEE Information Visualization 98, pp. 102-105, 1998.

[10] B.G. Becker, Volume rendering for relational data, in proceedings IEEE Information Visualization 97, pp. 87-91, 1997.

[11] Iris Explorer, http://www.nag.co.uk/Welcome_IEC.html

[12] Michael Friendly, Corrgrams: exploratory displays for correlation matrices, The American Statistician, vol 56, pp 316-324, 2002.

[13] Heijs and Smeulders, Visualization experiments of financial data using MineSet, to be published.

[14] Jonathan Roberts, Nadia Boukhelifa and Peter Rodgers. Multiform Glyph Based Search Result Visualization. In *Proceeding Information Visualization 2002*. IVS, IEEE, 549-554. July 2002.