IVC'05 Exploration Toolkit

Daniel Keim, Jörn Schneidewind, Christian Panse, Mike Sips, Jakob Haddick, Fabian Dill, Henrico Dolfing
University of Konstanz, Germany
{keim,schneide,panse,sips,haddick,dill,dolfing}@inf.uni-konstanz.de

1 Motivation

Providing effective and intuitive visualization tools for the InfoVis Contest 2005 tasks is a challenging ambition. Due to the size, complexity and data-type variety of the given dataset, most of the standard analysis techniques are inappropriate to any search for efficient solutions for the contest tasks. Since the dataset contains spatio-temporal, non-spatial and temporal data we need instead a combination of techniques which are specifically able to handle these different types of data. Therefore, to provide effective visualizations for the contest tasks, we created the IVC'05 Toolkit, a framework that contains a variety of visualization and interaction techniques for the exploration of the spatial, non-spatial and temporal patterns in the data . All visualizations and actions in the toolkit are tightly coupled using the well-known linking and brushing concepts.

2 Data Preprocessing

When processing and visualizing large data sets, data cleaning as part of data pre-processing is a very important but costly step, since it directly influences the quality of the visualization. In this context , the elimination of missing values was one of the main tasks. We replaced missing values by default values (e.g. average value) and in some cases by linear interpolation (i.e. number of sales, employees). To access and manage the given dataset efficiently we employed a PostgreSQL databases. The toolkit is implemented in Java, JDBC is used to extract the needed information from the database.

3 Overall Concept

To find solutions for the 3 contest tasks, we employed different visualization approaches. To identify correlations in the non-spatial temporal data, we applied the CircleView technique [Keim et al. 2004]. The basic idea of this approach is to divide an circle into segments according to the number of attributes and then to subdivided each segment according to a number of instances of each specific attribute.
To explore geo-related-patterns, we used a traditional US map , given as a set of polygons, and applied color, pixel graphics, histograms, distortion techniques and glyphs to visually represent relevant information. All techniques provide interaction functionality, like mouse over effects and visual query capabilities.

4 Analysis of the Contest data set

The basic idea of our toolkit is to provide some predefined techniques for the different tasks, like the CircleView technique for correlation analysis. The user than selects relevant attributes or regions of interest via interactive query sliders and buttons. Based on the chosen technique and the selected attributes our system generates an SQL query which will be send to the database. The returned result set will than be visualized by the selected technique.

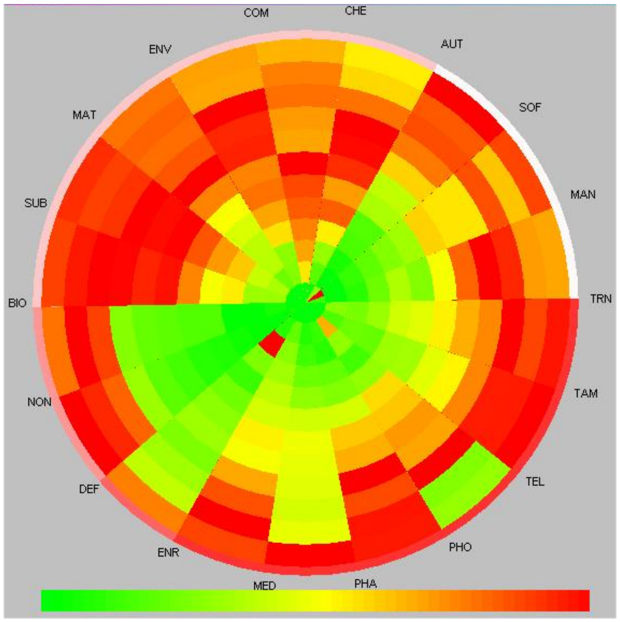

Figure 1: CircleView showing the number of sales per industry category from 1989 to 2003.

4.1 TASK 1: Characterize correlations or other patterns among two or more variables in the data

To identify correlations in the data, we employed the CircleView technique. Figure 1 shows the development of the total number of company sales per industry category. The segments at the center of the circle symbolize the sales amount for the year 1989, at the outside of the circle the sales amount for 2003 is visualized. The segments are clustered using the k-means algorithm. Clusters are visualized by small segments at the outside of the circle. An interesting information the user can extract from Figure 1 is that software (SOF) has an increase in number of sales in 2001 (yellow changed to orange), but a decrease in 2002 (orange changed to yellow). Additionally the clustering reveals that the total number of sales of software (SOF) and automotive industry (AUT) developed in an similar manner. Of course the user can interactively change the attributes for segment partition and color mapping, in order to explore correlations between any attributes of interest.

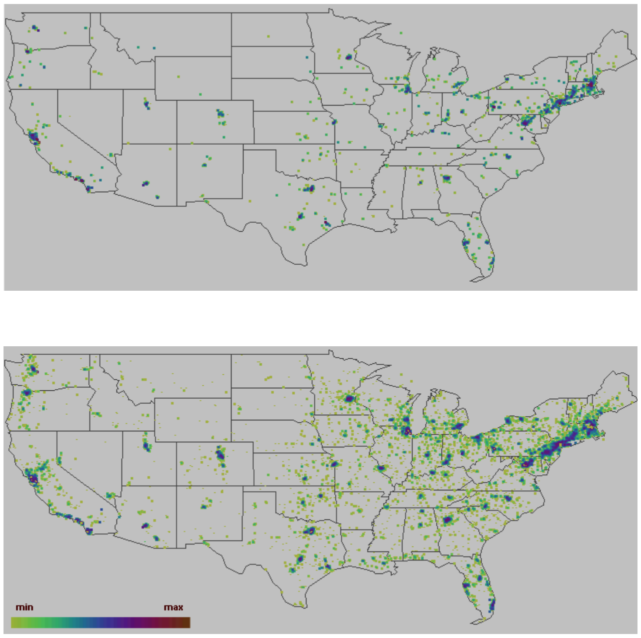

Figure 2: Dot Map showing development of industry from 1989 (upper figure) till 2003(lower figure). Color indications how many companies belong to each particular x/y location

4.2 TASK 2: Characterize clusters of products, industries, sales, regions, and/or companies.

For this task we employed a 2D US-map (without Alaska, Hawaii) and provided a number of methods for visualizing data on this map. This includes simple dot plots like shown in figure 2,Color Maps shown in figure 3 , Pixelmaps [Keim et al. 2003] and cartograms. Figure 2 clearly shows areas (States, Cities) with a high industry density (New York, California). The color indicates how many companies to each particular x/y location (zip code) belong, but the user can change the color attribute to any other existing attribute like number of products, employees etc. The upper figure shows the data from year 1989 and the lower one the data from 2003. Its easy to see that there is a strong increase in the number of companies.

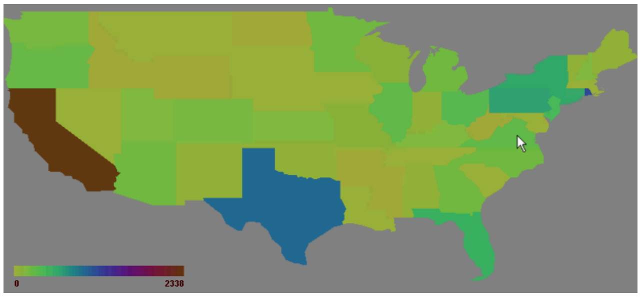

Figure 3: The figure shows the number of founded companies per State in 1989. Most companies where founded in California (2338).

4.3 TASK 3: Characterize unusual products, sales, regions, or companies

To detect unusual patterns and outliers, CircleView as well as our proposed spatial techniques are applicable. In figure 1 for example, the number of sales for the telecommunication (TEL) sector in 2002/2003 is unusual, since after a peak in 2001 it decreased clearly, while almost all other sectors increased their sales. Figure 4 investigates the movements of these companies, where the number of sales changed disproportionately after their movement.

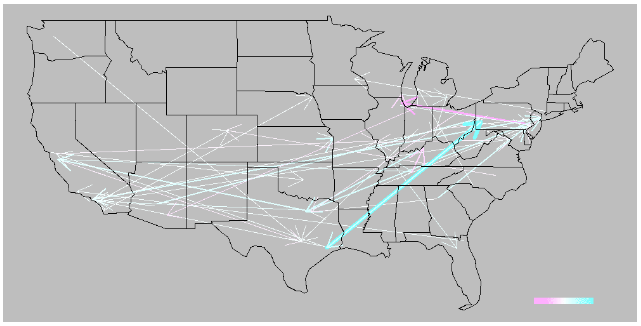

Figure 4: Visualizing the movements of companies from 1989 to 2003. To reduce occlusion only companies with 10 percent more / less sales after movement are shown. As shown many companies moved to California and New York. The green arrow shows a company that moved from Texas to New York accompanied by a maximum increase in number of sales.

Acknowledgement
Many thanks to Robert Spence for his valuable comments.

References

KEIM, D. A., PANSE, C., SIPS, M., AND NORTH, S. C. 2003. Pixelmaps: A new visual data mining approach for analyzing large spatial data sets. In *Proceedings of the 3rd IEEE International Conference on Data Mining, ICDM 2003, Melbourne,FL,USA*, 565–568.

KEIM, D. A., SCHNEIDEWIND, J., AND SIPS, M. 2004. CircleView: a new approach for visualizing time-related multidimensional data sets. In *Proceedings of the working conference on Advanced visual interfaces, AVI 2004, Gallipoli, Italy*, 179–182.