

Discovering USA technology trends with DARE and SpringView

Enrico Bertini, Luigi Dell'Aquila, Giuseppe Santucci

Dipartimento di Informatica e Sistemistica - Università di Roma "La Sapienza"

Via Salaria, 113 - 00198 Roma, Italy - {bertini, dellaquila, santucci}@dis.uniroma1.it

ABSTRACT

In this paper we discuss how the DARE and SpringView systems have been used to explore the InfoVis05 contest data set. We describe the environments we used, the data preprocessing, and the insights we gained with our systems. Moreover, we point out the pros and the cons of our approach and the lessons we learned.

1 INTRODUCTION

To explore the contest data set we used two general purpose visual environments, DARE (Drawing Adequate Representation) and SpringView, developed at University of Rome "La Sapienza". In order to deal with the tasks described in the form, we slightly modified the user interface, providing some shortcuts useful for the contest activities. The paper is structured as follows: Section 2 and Section 3 describe the DARE and the SpringView systems, respectively; Section 4 describes the data preprocessing and Section 5 deals with the insights we found in the dataset. Finally, Section 6 discusses pros and cons of our approach, pointing out the lessons we learned in this challenging contest.

2 THE DARE SYSTEM

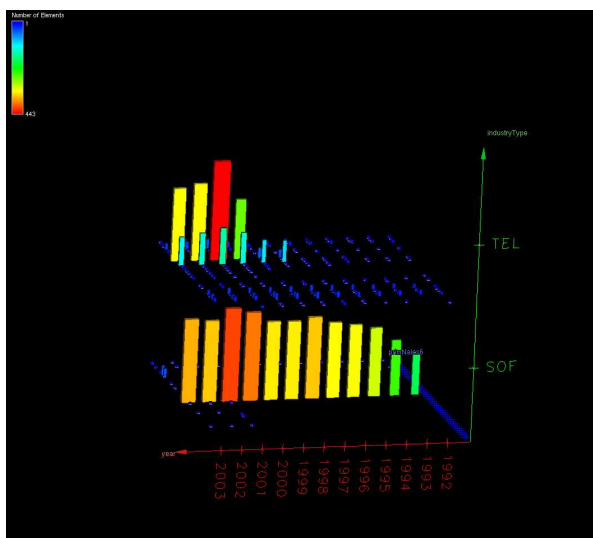


Figure 1: The DARE system:OLAP view

The DARE system [2] has been implemented to visually analyze large amounts of data, either exploring single data points' values or interacting with OLAP cubes to discover aggregate values. Data browsing is implemented using up to 6 visual attributes, i.e., x, y, z axes, color, size, and shape. The association between data values and visual attributes is performed manually or automatically, exploiting an ad hoc knowledge base. The OLAP visualization handles 1D, 2D, and 3D visual cubes showing, through the size and the

color of each cube element, two summary values (textual reports are available as well). Usual operations of drill down, roll up, and slice and dice are provided. Moreover, the user can switch between elementary and aggregate data at any time, visualizing different data set attributes.

In Figure 1 one of the contest activities is shown. The image presents a 3D OLAP data cube containing SOFTWARE and TELEcommunication companies grouped by year and primary NAICS; for readability purpose, both color and size denote the number of companies. It is quite evident that SOF companies with primary NAICS 511210 (SW publishers) grew between 1992 and 2003, while TEL companies started growing only in 1998, with primary NAICS 513330 (Telecommunication resellers) and much faster than with primary NAICS 514191 (On line information services); the number of On line information service companies in 2001 was greater than SOF companies. Both TEL and SOF companies slightly decreased in 2002 and 2003.

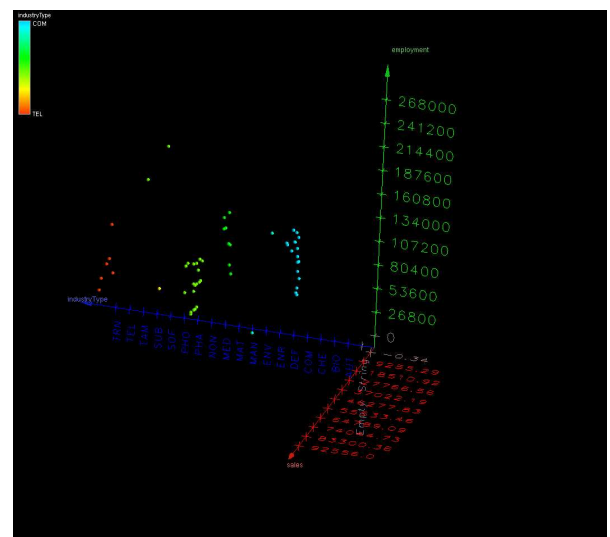


Figure 2: The DARE system:scatter plot view

Figure 2 shows some outliers in a 3D scatter plot: the images show companies with very high employment (26000 to 268000 employees) and sales counts (9200 to 92000 million dollars) in the observed years. Few items (66) are on the screen, all but one belonging to TELEcommunication, NON primary high-tech, and COMPUTER hardware company types. The isolated company is a SOFTWARE company (ID 7050, city Armonk, state NY) and its activity is Management of companies and enterprises (NAICS 55). The top-most two points correspond to the same SOF company (ID 72312) still with NAICS 55 that in 2002 had sales equal to 66565 million dollars with 268000 employees and that in 2003 increased its sales up to 92556 million dollars while losing (firing?) 8000 employees. Both color and Z axis represent industry type, helping the user in understanding ambiguous 2D projections of the 3D scatter plot, e.g., points belonging to NON companies are all green.

3 THE SPRINGVIEW SYSTEM

SpringView [1] has been specifically designed to deal with multi-dimensional data and integrate radviz [3] and parallel coordinates views exploiting their contrasting characteristics. From one side radviz offers good direct data manipulation (i.e., brushing) techniques and low cluttering but it fails in providing visualization of quantitative information; conversely, parallel coordinates clearly shows the values of data attributes and their ranges but suffers from high cluttering even with small datasets and presents tedious manipulation techniques.

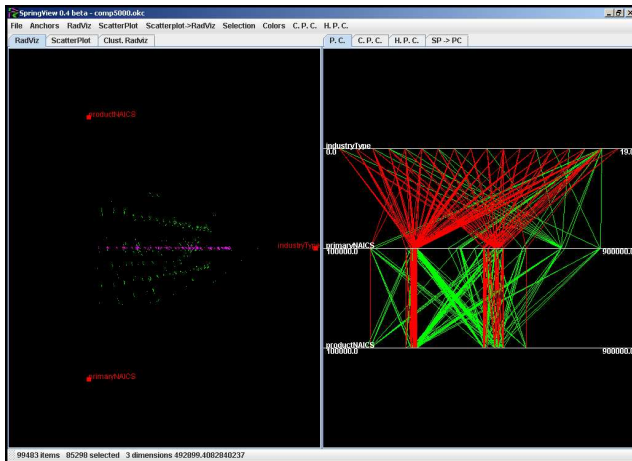


Figure 3: The SpringView system:n-Dimensional brushing on radviz

Figure 3 shows the usage of SpringView to quickly locate companies that dealt in some year with products *non* strictly related with their primary NAICS. The image displays a view of the whole data set, showing for each company the industry type, the primary NAICS and the NAICS of the products sold across the observed period of time. While the parallel coordinates view is quite crowded the radviz representation shows clear clusters. In particular, since primary NAICS and product NAICS are coded with numbers, the fact that a company X sold a product Y having the same NAICS as the company primary NAICS is depicted by a perfectly vertical line, which is clearly distinguishable from the others. It is possible to select all these points, highlighted in purple in figure, and brush the corresponding items on the parallel coordinates view (red items). The non selected items (green) represent companies that sold in some year at least a product having a NAICS code different from the company primary NAICS; such a data subset can be exported in the DARE environment for further analysis. It is worth noting that it is impossible to perform the above selection directly on the parallel coordinates view.

4 DATA PREPROCESSING

In order to answer the contest questions the original data set was quite heavily preprocessed. We computed joins among the data tables, and derived 10 new attributes, e.g., the age of a company, the mobility (in terms of the number of different locations in which the company has been), the number of different products a company produced in each year, the sales/employee ratio, etc. Moreover, we added the information about USA regions (Middle Atlantic, Mid West, New England, etc.) in order to exploit a deeper hierarchy on companies' location. Most of the results we obtained are based on the visual inspection of these values or their aggregation. The initial data exploration was performed on a sample of the whole data set, in order to speed up the system performance; once we had the main

trends we worked with the whole data set, to get precise figures and outliers.

5 INSIGHTS

Most of the data insights have been discovered using the 3D OLAP visualization of DARE. We used the other visualizations to have detailed information about data or to find some outliers. The tasks we performed are associated with all the three main contest questions' categories, dealing with: attribute correlations and other dependencies, clusters of similar data, both geographical and temporal trends, outliers.

6 BENEFITS, LIMITATIONS, AND LESSONS LEARNED

Perhaps one of the main strengths of our approach is the availability of different visual representations working on the same data set at the same time and the possibility to easily focus on interesting subsets for further investigation. The possibility of working with both aggregate and not-aggregate views proved particularly useful.

Speaking of limitations, we noted that some operations can be quite annoying, e.g., browsing year by year some aggregate values can be difficult since many data points appear and disappear abruptly. Set up time is also a matter, obtaining the visualization one has in mind is not always easy; the typical pattern we experienced was to obtain an initial view and then correct it according to the task one has in mind. Moreover, data preprocessing was really time consuming and difficult to perform.

Concerning lessons learned we have several issues:

- Visual attribute overloading can be very useful. We often used two visual attributes (e.g., bar height and color) to represent the same data dimension. This can drastically increase the effectiveness of a representation especially when disambiguation of 3D objects is needed.
- Text (and paper!) is useful. While 3D visual OLAP cube are quite effective to quickly spot trends and other interesting features, there always is a strong need for labels, tables, and textual reports to compare the values in the display; having in DARE both textual and visual representations was a useful feature. Moreover, we often found ourselves using printed documents to compare values and to make sense of data. We believe it is important to recognize that visual exploration cannot only happen by looking at data on the screen, paper is still the best tool for many purposes.
- Data preprocessing is crucial and should not be underestimated. It deserves more attention since it is difficult to perform, time consuming, and crucial to obtain effective visualizations. Most of the insights we had come from derived or added attributes. We believe this is a general rule: effective visualizations often come from an intense data preprocessing.

REFERENCES

- [1] Enrico Bertini, Luigi Dell'Aquila, and Giuseppe Santucci. Springview: Cooperation of radviz and parallel coordinates for view optimization and clutter reduction. In *Proc. of IEEE CMV 2005*, pages 22–29, 2005.
- [2] Tiziana Catarci and Giuseppe Santucci. The prototype of the dare system. In *ACM Proc. of SIGMOD International Conference on Management of Data*, 2001.
- [3] Patrick Hoffman, Georges Grinstein, and David Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *NPIVM '99: Proc. Workshop on New Paradigms in Information Visualization and Manipulation*, pages 9–16. ACM Press, 1999.