Please see supplementary material on conference DVD.

SecureScope™ Visualizations of the NAICS Dataset

Brianne O'Brien, Victor Seguritan, Dan Tesone, Anita D'Amico, Ph.D.

Secure Decisions, a division of Applied Visions, Inc.[1]

ABSTRACT

This summary describes an analysis of the NAICS dataset using SecureScope, a three-dimensional visualization tool for exploring relational data. SecureScope was developed by Secure Decisions under a Small Business Innovative Research (SBIR) grant from the US Air Force, and enhanced under an SBIR grant from the Defense Advanced Research Projects Agency (DARPA). SecureScope was originally designed to analyze data related to information security but can be extended, as illustrated here, to analyze many kinds of relational data. SecureScope retrieves and clusters data, creates associations between SecureScope objects representing data points or groups, and presents the results as an interactive, three-dimensional graphical scene to aid users in the interpretation of large datasets and to improve their understanding and awareness of the data being analyzed.

1. ANALYSIS PROCESS

The NAICS dataset was first imported into an Oracle 9i database. SecureScope was then "taught" the schema of the NAICS database through the use of a metadata generation utility program developed specifically for SecureScope. The COMPANYDATA table was updated to include a field called SALES_EMP, and a simple script was written to populate the database field with calculated sales-per-employee ratios obtained from data within the COMPANYDATA table. Additional "display property" configurations were generated which defined the color, size, blink, and motion characteristics of each data field visualized within a scene. Contest questions were answered using SecureScope's Grid and Wall scenes, along with statistical analysis using NAICS data retrieved from the Oracle database. An example of how visualized results and statistical profiles are typically rendered by SecureScope is presented in Figure 1:

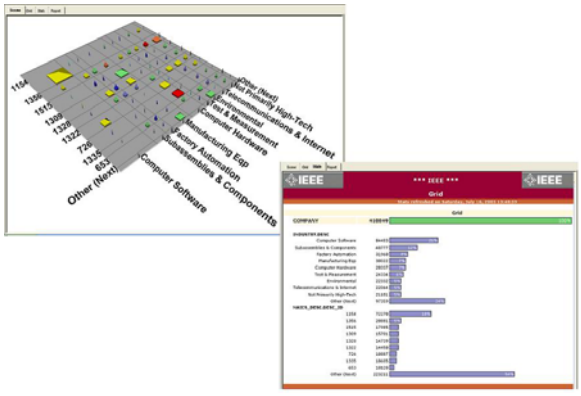

Figure 1 – SecureScope scene and sample statistical analysis

1.1 Characterization of Correlations and Patterns

Question 1, regarding the characterization of correlations and patterns among two or more variables in the data, was addressed using SecureScope's *Frequency Wall* scene. A Frequency Wall is a temporal visualization containing a vertical grid displaying time points on the top horizontal axis, and a database field selected by the user on the vertical axis. The time point selected was the YEAR field from the COMPANYDATA table, and the INDUSTRY_DESC field from the INDUSTRYCODES table was chosen to be displayed on the vertical axis. Individual bin items contain bar graphs (histogrids) which are visually sized according to a count of Companies within the COMPANYDATA table. They are sized relative to each other within the scene.

SecureScope provides the capability to drill-in to any individual bar graph so that more information can be obtained for the selected item. The bar graphs are colored according to the sales revenues associated with each group of companies, and groups with particularly high sales are configured to *blink*. This scene illustrates the growth, followed by decline, of several high-tech industries. It also demonstrates the decline, followed by growth, of *other* industries that are defined as "Not High-Tech" in the database.

Further confirmation of the trends revealed by the first scene is provided by a similar SecureScope *Grid* scene. While the first visualization demonstrates growth and decline in terms of company numbers and sales revenue, this supplemental visualization displays similar trends based upon the number of employees within each industry over time. This visualization is more demonstrative in showing industry growth and decline trends, by looking at workforce participation in various industries rather than at sales revenue. A reinvigoration of the workforce within the non-high tech industry can be clearly seen in 2001 and, conversely, the decline in high-tech industries – particularly in the Telecommunications & Internet sectors – is reflective of the popping of the "dot com" bubble.

1.2 Characterization of Clusters

Question 2, regarding the characterization of clusters of products, industries, sales, regions and/or companies, was addressed using SecureScope's *Grid* scene. The Grid view provides a visualization displaying relationships amongst the clustered items on the grid.

The items contained within each Grid bin, represented by different-sized pyramids, are derived from a user-selectable database table and represent the number of items within that table sorted by the properties selected for the X and Z axes. The Grid represents an overall "big picture" view of the number of companies, organized by industry and NAICS code. The larger the pyramid, the more companies are contained within that cluster.

Every pyramid is colored according to the maximum range of revenue for a company within a cluster. With this visualization, the number of companies and maximum revenue of companies within an industry are apparent.

Pyramid size and color are particularly striking in this visualization, and clearly distinguish several different industries with regard to their very large presence and high sales. Even more striking are the *smaller* pyramids, indicating fewer companies

[1]6 Bayview Avenue, Northport, NY 11768
BrianneO@SecureDecisions.com

with colors reflecting higher sales. Drilling into a selected pyramid provides details for the group of companies represented within that pyramid.

The Computer Hardware and Computer Software industries are of particular interest. By drilling in, new Grid-type visualizations are generated that reflect a distribution of companies by state and year within each computer industry. The resulting visualization for the Hardware industry demonstrates definitive strength in the California region in terms of company numbers and sales revenue, particularly in 1993. We can theorize that this surge reflects the thrust towards the goal of having a PC on every desk in the workplace, and every room in the home, by the end of the millennium.

1.3 Characterization of unusual items

Question 3, regarding characterization of unusual products, sales, regions and/or companies, was also addressed using SecureScope's *Grid* scene. Product trends by region and state are shown, with product year and company state chosen as axis properties. Company sales in millions of dollars are color-coded in the same manner as described for Question 2.

To demonstrate the value of visualizing a specific industry, NAICS data was selected for companies in the Pharmaceutical industry. By focusing on one industry, the effect of a region's growth on neighboring regions for a specific industry is apparent. The top revenue and product producers displayed for this visualization are in accord with the ten major US regions defined in Biospace.com, a web-based resource for the life sciences industry. The size and color of pharmaceutical clusters in NJ, CA, PA, NY, and MA are the most prominent in this display. Other notable clusters are seen in mid-Western and Southern states.

Surprisingly, this SecureScope scene discovered a steady increase in product growth and sales for the two "BioSoutheast" states of Kentucky and Tennessee, starting in 2001. The pattern of growth seen in Kentucky and Tennessee parallels the growth of neighboring states, such as Illinois, Indiana, and North Carolina. These surrounding states with a large number of high-revenue pharmaceutical companies may be providing business to states within close proximity, such as Kentucky and Tennessee.

Another unusual pattern in the NAICS data is demonstrated by the configuration of groups of companies organized according to industry description and state, focusing on the "top 10" states in terms of sales. Of particular prominence is the small number of Arkansas-based companies in the Telecommunications & Internet sector responsible for enormous sales during 1993. Two companies responsible for these large sales revenues were discovered using just two drill-in visualizations on the Arkansas cluster: for one drill-in scene, state and industry descriptions were chosen as axis properties; a second drill-in scene was created using state and year data from the COMPANYDATA table. The two drill-in visualizations uncovered these Arkansas-based companies with unparalleled sales in 1993.

1.4 Characterization of any other trends

Question 4, regarding the characterization of any other trends, was addressed using SecureScope's *Wall* scene. This scene was generated based on groups of companies, and organized according to their industry. *Year* is color-coded according to sales-per-employee figures. The SecureScope Wall scene using the sales-per-employee ratio (SPE) provides a visualization that enables easy comparison of the productivity of companies within an industry to other industries over time. Cyclical trends in SPE were predominantly seen in a few industries, while some industries displayed trends that were parallel to other industries.

Specific industries such as Advanced Materials and Factory Automation are distinguished from others for extremely high sales-per-employee figures. We can theorize that the SPE for the Factory Automation industry, for example, peaked in 2003 as a consequence of the productivity in industries upon which Factory Automation strongly depends, such as Computer Hardware, Computer Software, Subassemblies & Components, and Telecommunications & Internet.

Of particular interest, demonstrated well by color-coding, is the decline and then subsequent recovery of the Telecommunication & Internet industry by 2003. In addition, the growth and decline of the Non-High-Tech and High-Tech industries appear to be mutually exclusive, and are displayed well using the parameters configured for this visualization, as shown in Figure 2.

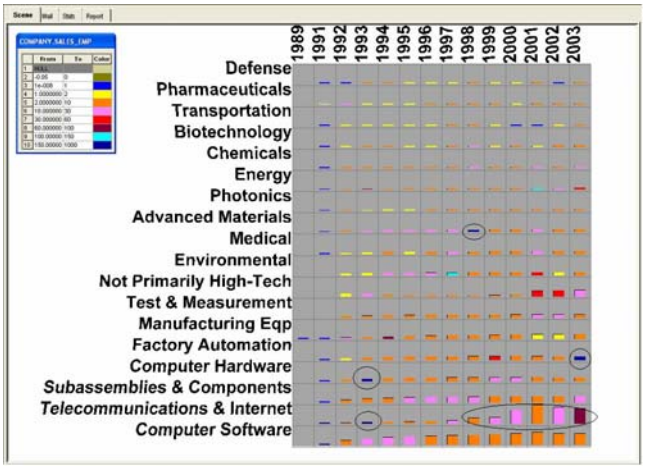

Figure 2 – SecureScope Wall scene

2. STRENGTHS

SecureScope is capable of visualizing data and data associations in a single graphical scene, using a number of visual cues. All SecureScope scenes are configurable to user specifications, making this tool generic enough to use on any type of domain data contained within a relational database. The selections of visualization templates are designed to support different types of data analysis, providing ease of use for temporal analysis, analysis of associations among large datasets, and mission- and business-impact analysis.

Drill-in features that provide access to more-detailed scenes, or even to the underlying raw data supporting a visualized scene, together with statistical analysis of each scene, provide for multiple methods of data analysis and reporting.

The uses of color, motion, blink, size and other visual cues allow the user to easily and quickly identify extraordinary or unusual items and trends in large datasets.

3. WEAKNESSES

SecureScope was originally developed to visualize a large dataset of network intrusion data, and has matured to where it can be used to visualize many other types of data. SecureScope does not, however, include the ability to apply complex mathematical operations or algorithms to the data that is being visualized. A forecast analysis using yearly sales data by industry, for example, would be a good statistical measure for sales trends across different industries. In addition, a *link analysis* form of visualization may be helpful in determining the accelerator principle impact of decline and growth across and within industries.

Such visualizations can be helpful in seeing more "intimate" relationships in data, and the more subtle patterns produced by slower-moving forces in the economy.