ABSTRACT

SAS® Visual Analytics Explorer is an advanced data visualization and exploratory data analysis application that is a component of the SAS Visual Analytics solution. It excels at handling big data problems like the VAST challenge. With a wide range of visual analytics features and the ability to scale to massive datasets, SAS Visual Analytics Explorer enables analysts to find patterns and relationships quickly and easily, no matter the size of their data. In this summary paper, we explain how we used SAS Visual Analytics Explorer to solve the VAST Challenge 2012 mini-challenge 1.

Keywords: Visual analytics, big data, data visualization, exploratory data analysis.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces – (Graphical User Interfaces); I.3.6 [Computer Graphics]: Methodology and Techniques – (Interaction Techniques);

1 INTRODUCTION

SAS Visual Analytics Explorer [1] is a Web-based exploratory data analysis application designed to enable the discovery of patterns and anomalies in large volumes of data [2].

In this year’s VAST Challenge mini-challenge 1, participants were asked to analyze 158 million log entries from a fictitious bank’s computer network in order to assess the overall status (situation awareness) and to find up to 5 anomalies. Handling a dataset of this scale is usually a challenge in itself. SAS Visual Analytics Explorer uses the SAS LASR Analytic Server to address scalability issues related to analyzing large volumes of data. SAS LASR Analytic Server holds entire datasets in memory and evaluates queries by using parallel processing algorithms, thus enabling SAS Visual Analytics Explorer to create complex visualizations in seconds. This level of performance gives analysts the ability to maintain an uninterrupted workflow, cycling quickly between hypothesis generation and verification. It also provides analytics, data filters, dynamic hierarchies, and interactive brushing and linking to connect multiple visualizations.

In the next sections, we describe how these features were used in the resolution of the VAST Challenge.

2 FEATURE OVERVIEW

2.1 Data Visualization with Analytical Support

Answering the situational awareness question required aggregated visualizations that still preserved the “big picture.” We created heat maps based on the latitude and longitude of each log entry for the specified timestamp. For the color response, we relied on descriptive analytic functions of the measures that we were investigating. Figure 1 illustrates how branch working hours, displaced by their respective time zones, were reflected in the geographical distribution of the average activity.

Figure 1: Heat map of BankWorld’s activity

2.2 Data Filtering

Next, we looked for anomalies across the entire dataset. We used data filtering to isolate them as needed. Figure 2 shows how the number of online machines seems to increase during the second day when compared to the same hour on the first day.

Figure 2: Count of log entries over time for all machines
However, this visualization does not provide information on the specific reason behind this pattern. We knew that ATM and server machines are always online. We also knew that workstation activity is governed by the branch’s working hours. We had to separate these multiple, conflicting patterns to make sense of the anomaly.

We created a filter based on machine class to view each class independently. The filter revealed a difference of almost 50,000 servers from one day to the next, as shown in Figure 3.

2.3 Hierarchies and Drilling

SAS Visual Analytics Explorer enabled us to create categorical hierarchies without the need for pre-aggregation. Hierarchies can be used in tree maps [3] and other visualizations to explore the data through drill and expand operations. Consider the following hierarchy:

```
BusinessUnit -> BusinessFacility -> MachineClass -> MachineFunction -> IPAddress
```

We used this hierarchy to analyze patterns of virus infection across the network, from entire units down to individual machines, as illustrated by Figure 4. Here, we see multiple tree maps in progressively deeper drilling states. Because infected machines are indicated by the highest policy status value, we associated the color response with the maximum aggregation for this measure. The result was that the color red indicated the presence of at least one infected machine in the respective area. At a glance, we can tell that the infection is widespread at a high level. However, as we drill down, we find fewer and fewer occurrences. This distribution implies a sparse infection pattern.

Data brushing and linking allowed us to interact with the data and look for patterns that we could later isolate by using filters. Figure 5 shows how we used a horizontal lattice of frequency plots (right) to isolate the difference in the number of logs across machine functions per day. This lattice was then used as a starting point to find out the corresponding values in linked frequency plots of latitude and longitude (left). This process allowed us to identify geographical patterns in the missing values, which we later were able to associate with one specific bank region.

2.4 Data Brushing and Linking

Although filtering was essential for isolating anomalies, it was not always evident which columns and values needed to be filtered.

3 CONCLUSION

SAS Visual Analytics Explorer is the latest member of the SAS Institute extensive family of data analysis and visualization applications and tools [4, 5].

The scalability, analytics, and advanced data visualization features of the SAS Visual Analytics Explorer proved to be an excellent match for the demands of VAST Challenge 2012.

REFERENCES


