VAST Challenge 2012: Interactively Finding Anomalies in Geo-temporal Multivariate Data

Gosia Migut* Justin van Wees Diederik Bakker Bart de Goede Haska Steltenphol Nick Oude Lenferink
Marcel Worring

University of Amsterdam
The Netherlands

ABSTRACT

In this paper we describe the tool we developed to solve the VAST 2012 Challenge. We present how an expert can online analyze huge amounts of multivariate data in space and time. The visual components and the interaction between the system and the user are described, as well as the setup to allow on the fly processing and retrieval of huge amounts of data. We summarize how our tool helped us to solve the challenge.

Keywords: visual analytics, interactive visualization, intelligence analysis, human information interaction

1 INTRODUCTION

The VAST 2012 Challenge introduces the problem of analyzing the Bank of Money network containing nearly a million machines with over a hundred status updates over a period of two days. Our objective in developing a tool to solve the VAST 2012 Challenge is the ability to support scenario generation processes for situation awareness for huge amounts of data. Athough our system is specifically aimed at analyzing the Bank of World data, it can be used for other situation awareness problems where relations between geo, temporal and multivariate variables have to be discovered.

To deal with a million of entries, we have to make a number of choices. The data has to be either analyzed first and the precomputed set of data visualized, or the data can be analyzed and visualized on the fly according to the analyst's needs. Since we consider user interaction as the key component in effectively analyzing a domain specific problem, we see option two as the most effective solution. For processing huge amount of data on the fly, including storing the data and retrieving it in a fast manner based on the user queries, we need to find a smart solution. Having such an architecture at our disposal, our second concern is allowing the user to interactively find the relation between geo, temporal and multivariate data.

In the next section, we describe in detail how the framework is set up our choice of appopriate visualizations to allow efficient scenario generation in situation awareness.

2 ANALYTIC PROCESS

To be able to discover which anomalies arise in the Bank of Money computer network during the two day time period, the architecture has to meet several requirements. With more than 158 million status messages from approximately 895,000 different machines spread across 4,056 different locations, allowing the user to view different aggregations of the data is essential. In addition to the geographical characteristics the diachronical aspect of the data plays an important role in discovering patterns and trends. This means that the user has to be able to get an overview of the events taking place over time on various aggregation levels, and in some cases needs the ability to 'zoom-in' on a specific point in time. A very important and also challenging requirement is to make the system sufficiently 'responsive' to the queries resulting from the interactive exploration of the user.

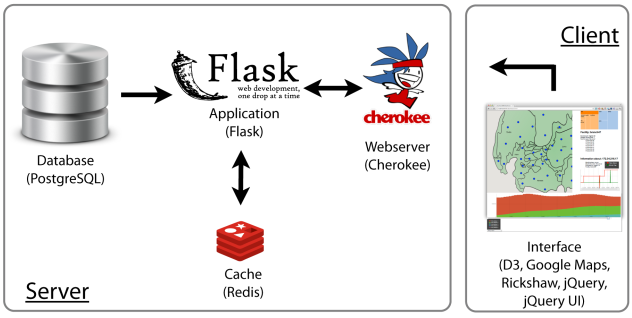

Figure 1: A server-client architecture for on-the-fly query processing.

2.1 Architecture

Libraries such as D3.js [1] make it possible to make responsive visualizations within modern browsers. To prevent the browser from having to parse several gigabytes of data (which would be far to slow), we opt for a – very common – client-server model. Storing, processing and caching of data are done on the server, visualization and manipulation of the data on the client.

The data provided by the VAST Challenge is stored in a PostgreSQL [1]database. Indexes on different (combinations of) data fields are constructed in order to speed up retrieval of specific parts of the data. We wrote software in the Python-based microframework Flask [2] to retrieve data from the database, perform aggregations and to handle client-side requests. Redis [3], an in-memory key-value data store, is used as a caching layer between the database and our Flask application. Among other things, Redis stores aggregations in memory for faster retrieval. Figure 1 shows a schematic overview of our architecture.

As mentioned, D3.js is used as a browser-based visualization toolkit. Google Maps [4] with a custom layer is used for the geo-

*corresponding author: mmigut@gmail.com

IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

[1]http://www.postgresql.org/
[2]http://flask.pocoo.org/
[3]http://redis.io/
[4]https://developers.google.com/maps/

visualization, the jQuery [5] framework helps in handling interaction and communication with the server. Data between client and server is exchanged in JSON.

2.2 Visualization framework

To keep the tool simple and clear we choose to follow Schneiderman's mantra: "overview, zoom & filter, details-on-demand". Since the dataset is low dimensional we use standard visual representations: a map for geo-data, a timeline for temporal data and a barchart for ordinal data. The interface is presented in figure 2. The visual components are interactively connected to reveal the relations between the variables of different type.

We use a map-centered view, to provide the analyst with the overview of the regions of the Bank of Money. Characteristics of each region are visualized as dots, where size and color intensity indicate the aggrageted value of the user selected numeric/ordinal variable (number of connections, policy status, policy flag). The choice has been made to aggregate the data over region in order to make visualization comprahensible for the analyst. If an unexpected value appears, analyst can zoom into the specific region to see what is happening.

The temporal information is encoded through the timeline, interactively connected to the geo-visualization i.e. the overview on the map is shown for each specific time-step selected by the user in the timeline. This allows to observe geographical changes over time. The temporal information is also encoded on the details-on-demand level. To zoom in into each aggregated region for details a simple barchart is used showing the trend of the aggregated numeric/ordinal variable for the specific region over time.

3 Finding anomalies in the Bank of Money network

In this section we demonstrate how we analyzed the challenge data with our tool. We report only the few observed anomalies that best illustrate the usefullness of our tool. The combination of the map-centered view interactively connected to the timeline, allowed us to discover most of the Bank of Money possible problems. Through the overview the major problems were identified. Then zooming in onto specific regions, allowed us to find suspicious patterns.

Firstly, we explored whether a suspicious policy status 5 showed up on a map by exploring each 15-minutes time step on the time line. Thus, we located a possible virus infection/questionable set of files found originating from datacenter 2. The first case of policy status 5 took place at 11.45 a.m. BMT on the 2nd of February. By zooming into this location we saw the trend of policy status 5 over time for this specific region. We observed that the amount of machines reporting status 5 in this region increases over time. Our analysis showed that the patient zero machine was 172.2.194.20, a server with a compute function.

Secondly, we noticed that machines in region 10 and 5 never report policy status 1. We discovered this by exploring the geo-visualization and selecting policy status 1 as node color variable. In this way, we could see the relative amount of machines reporting policy status 1 for each time-step as explored through the timeline. This functionality made regions 10 and 5 pop out on a map. A further analysis of the policy status 1 over time for region 5 and 10 supported our hypothesis, as it showed that no policy status 1 was never reported.

Thirdly, we noticed a large amount of machines stopped reporting in region 25. To explore the normal activity of the machines through time we looked at the number of machines reporting their status. Exploring the number of connections barchart for each region, we noticed that most regions show a similar pattern. By displaying activity flag 1 as the node color we noticed an anomaly on the geomap. When graphing the policy status reports over time, we

[5] http://jquery.com/

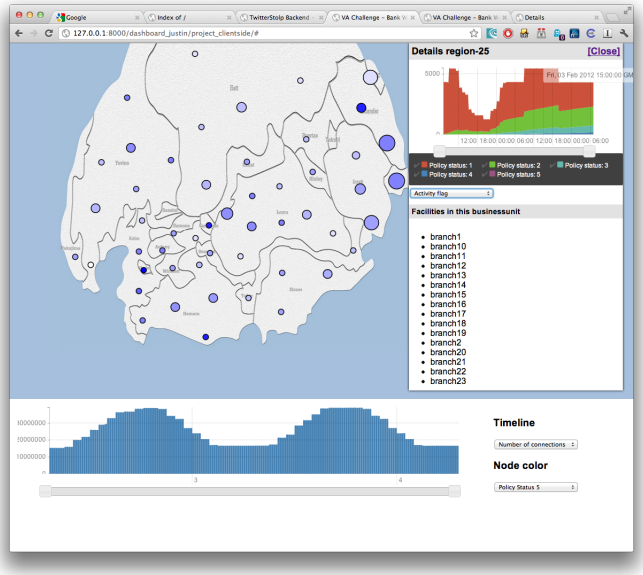

Figure 2: Policy status for region 25 over time.

noted a gap in machine activity for region 25 starting at 11.15 on the 2nd of February. This is illustrated in figure 2. The gap starts forming slowly and takes a sharp descend at 12.15p.m. Around 04.30 a.m. on the 3rd of February the anomaly is gone and the regular pattern found in other regions re-occurs in region 25. An interesting finding is that the downtime does not seem related to maintenance, as confirmed by the barchart of the activity flag 2 for this region. The downtime only occurs in certain branches of the region with no discernible pattern. Possible explanations could be an electricity blackout, strike, or natural disaster.

4 Discussion

A primary task of the VAST Challenge is to detect suspicious events in the Bank of Money network. The tool we developed helped us to identify the events that are potentially suspicious. The tool provides primarily an overview of health status over time for the entire network. To allow better analysis on the detailed level the tool could be extended with the functionality of zooming in into regions of interest, where the details of the particular groups of machines, or the particular machine over time could be visualized.

Acknowledgements

The authors wish to thank Denise Agathocleous, Tycho Bismeijer, Folkert Heeneman, Marije Meijer from the University of Amsterdam for providing static visualizations answering the MC1.1 question of the Challenge (see accompanying video).

References

[1] M. Bostock, V. Ogievetsky, and J. Heer. D3; data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301 –2309, dec. 2011.