OWLAP - using OLAP approach in anomaly detection

Award: Good Support for the Data Preparation, Analysis, and Presentation Process

L. Dudás Zs. Fekete J. Göbölös-Szabó A. Radnai Á. Salánki A. Szabó G. Szűcs
Computer and Automation Research Institute (MTA SZTAKI), Hungarian Academy of Sciences*
{ldudas, zsfekete, gszj, aradnai, salankia, aszabo, szgabbor}@ilab.sztaki.hu

ABSTRACT

OWLAP (Operative Workbench for Large-scale Analytics and Presentation) is a visual analytics tool that allows the user to browse and drill down the multidimensional data on-line with the possibility to export result into a zooming presentation framework. We address the challenges of multidimensional visualization by aiding the cognitively hard task of understanding attributes, finding patterns and outliers. We successfully solved the challenge of real time Big Data OLAP reporting by a home developed multithreaded in-memory database manager. Our additional focus is the automatic management of summary preparation that we aid by scripting the presentation framework of Prezi Inc.

Index Terms: H.2.1 [Information Systems]: Database Management—Logical Design; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 INTRODUCTION

We present our visualization tool for the VAST Mini Challenge 1 to analyze the "Bank of Money" (BOM) network log dataset. The goal is to find anomalies in the network where each computer has a series of log data of various actions and the current policy status within a 2-days time window.

The main challenge was to efficiently handle the size of the data: the BOM network contains approximately 900 000 computers with 192 temporal features describing them. The raw data amounted to near 4 GB. We reduced the data to circa 0.5 GB after implementing our own data structure, which gave us the opportunity to store all information in memory.

Each computer in the network has the following properties: type, function, region, facility and three time series. The three time series contain the policy status, activity performed lately and number of alive connections in the previous 15 minutes time interval.

Our aim was to develop a tool that can support the analyst in browsing and discovering the data while getting visual feedback about the data slice analyzed at the moment. More precisely, we merged the idea of OLAP (Online Analytical Processing) with a visual analytics tool [4]. Although there are existing software products for this use (Tableau Desktop [3] or RapidMiner [2]), we tailored our tool for this challenge, especially for anomaly detection.

Our solution is optimized for the VAST Challenge by zoom effects over the plot matrix. We may represent five of the dimensions of the data by the x and y coordinates, row, column and stacking. In our system OLAP queries can be formed and visualized over the geographic map by a drag-and-drop facility.

Another key feature of OWLAP is to provide for an analyst an automated way to generate a presentation about the steps and ideas that had come up during the analysis. For the presentation we used Prezi [1], which is a perfect tool to present the train of thought of the OLAP analysis.

*This work was supported by the EU FP7 SEC project SCIIMS (Ref. 218223) and OTKA CNK 77782

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

2 THE OLAP-BASED SOLUTION

Online Analytical Processing [5] has three main operations: *roll-up*, *drill-down* and *slicing-and-dicing*. In the table representing OLAP results we place stacked histograms and thus we can show five dimensions (x, y, row, column, stack) and aid in spotting charts deviating from the rest. Selected charts can be further explored by OLAP operations:

- *Roll-up* is performed when the user has already narrowed down the examined data using some filtering conditions. This operation is carried out by removing some of these filters.
- *Drill-down* happens by specifying dimensions and an aggregate function. The result is a plot matrix, whose items correspond to data subcubes that are results of the separation.
- *Slicing and dicing* is like zooming on a graph in the plot matrix (it can be also performed by defining filters by hand) to select a sub-datacube for analysis.

3 DIMENSIONS

Our graph matrix is able to handle up to five dimensions simultaneously. Drilling-down allows the user to separate the data by two dimensions, the individual graphs have configurable x and y axes, and the user can use different colors to split the data by one more dimension. Interactive pop-ups maintain the focus of the analyst when selecting the active dimensions. Examples for multidimensional visualization are shown in Fig. 2 and Fig. 1, left.

The analyst is not restricted to the given eight dimensions, she can define new ones by writing a short Java-style code. With this feature we produced dimensions like "policy status transitions", "external device attached to this computer". This flexibility is vital in any scenario when the anomalies are more complex and hidden deeply in the data, or the task is to develop an understanding of a new problem. User-defined dimensions can be used with the same ease as the default ones.

4 GEOSPATIAL VISUALIZATION

To investigate the geospatial characteristics of the observed anomalies, it was important to visualize the observed deviance on a map. After selecting a sub-datacube by OLAP-operations or simple filtering, one can move it onto the map. A small, colored circle belongs to each facility with radius proportional to the amount of computers that belong to the selected datacube and located in the respective facility. A color to each data set must be selected by the user. If several data cubes are visualized simultaneously on the map, then the color of a facility is blended by the number of computers that belong to the selected datasets. (See Figure 1).

5 PREZI EXPORT

After discovering deviant behavior, the analyst has to present the results to a non-expert (e.g. the CEO), therefore we provide the opportunity to the analyst to generate a presentation automatically. During the analysis the user can mark the most important graphs on the screen. Later these can be added to a Prezi presentation together with the filtering OLAP-information with one click. Technically, an XML file will be generated that describes the content and layout of the presentaion.

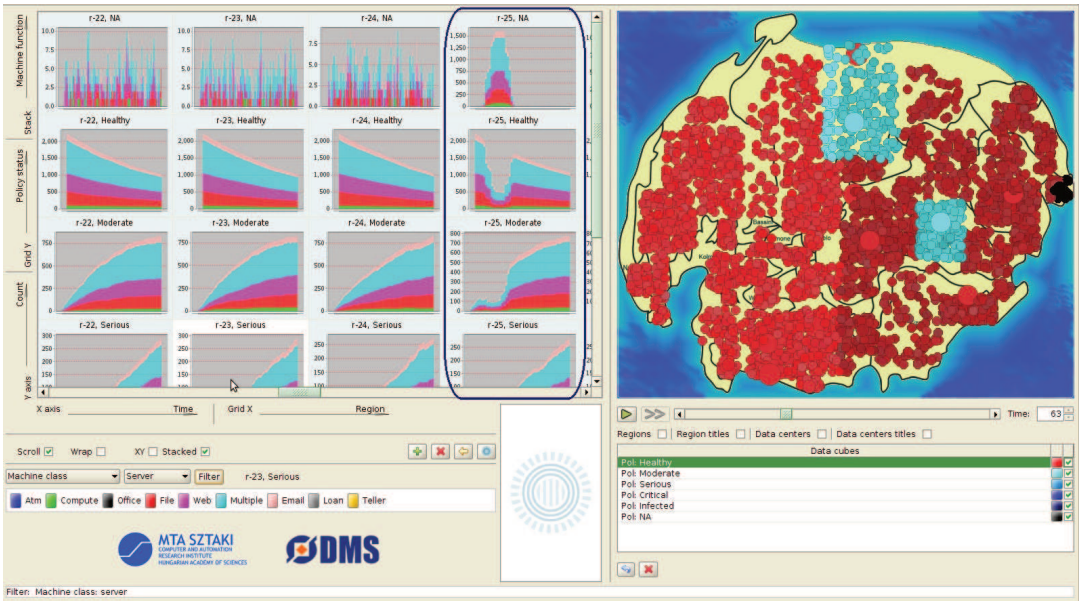

Figure 1: Usage of the map: Based on policy status one can define six sets. To observe the geospatial characteristics of these sets, the analyst can choose colors for them and create an animation. In this example two key anomalies can be found on the map: region 25 is black (because of a black-out) while region 5 and 10 contain only infected machines. Using the slider, one can also observe the trends: the change of days and nights, the direction of region-25 blackout propagation or the phenomenon that computers are getting more and more infected (the circles turn from red to purple) are easily detectable.

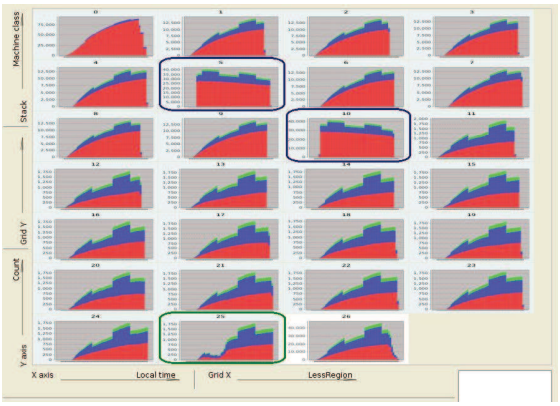

Figure 2: Anomaly detection with plot matrix. We drilled-down the data only by region (as column). Even if no row is selected, the graphs can be organized into matrix form. It is easy to recognize that charts 5, 10 and 25 do not follow the trend. On this screenshot four dimensions can be observed simultaneously: each graph represents a region and they show the number of moderately infected computers in time. Colors denote the different computer types (server, workstation and ATM).

The exported pictures will be arranged in a reasonable, hierarchical way on the Prezi's plane, and the sizes of the pictures will correspond to the level within the OLAP hierarchy (see Figure 3). This layout lets the CEO see the overall picture and relations between the steps of analysis, and, while „playing" the Prezi, nice pan and zoom effects will help reconstructing the analyst's train of thought. Information about the filters applied will be added to each picture as a caption, so even the details will be easy to follow.

6 Architecture

Since OWLAP has a server-client architecture, heavy computations can be performed while the graphical client runs on a simple personal computer. As OLAP operations are efficiently parallelizable,

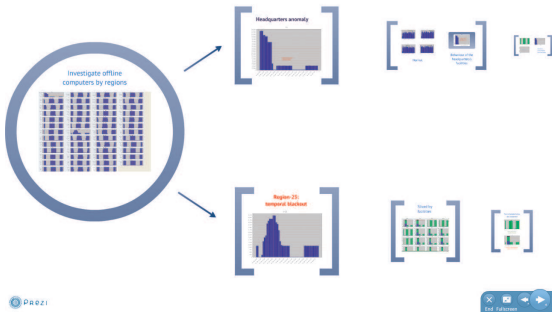

Figure 3: Overview of an exported prezi explaining two of the discovered anomalies.

the server generates OLAP cubes using eight parallel threads, dividing the data into smaller parts and each thread handles its own fragment. This way for any OLAP cube request the result is obtained within 10-20 seconds. The light-weight client is only responsible for the visualization.

References

[1] Prezi homepage. `http://www.prezi.com`. [Accessed Aug. 7, 2012].

[2] Rapidminer project. `http://sourceforge.net/projects/rapidminer/`. [Accessed Aug. 7, 2012].

[3] Tableau homeapage. `http://www.tableausoftware.com`. [Accessed Aug. 7, 2012].

[4] A. Cuzzocrea and S. Mansmann. Olap visualization: Models, issues, and techniques. *Encyclopedia of Data Warehousing and Mining*, pages 1439–1446, 2009.

[5] J. Han and M. Kamber. *Data Mining: Concepts and Techniques* . Morgan Kaufmann, 1st edition, 2000.