Simon Fraser University (SFU) - Tableau
VAST 2010 Mini Challenge 2

Minoo Erfani Joorabchi, Mona Erfani Joorabchi, Chris D Shaw

School of Interactive Arts and Technology, Simon Fraser University

ABSTRACT

For the VAST 2010 Mini Challenge 2, a commercial visual analysis system called Tableau was used to analyze and summarize the spread of the Drafa virus (disease) within each of the eleven countries, compare the occurrence time of the outbreak and identify the anomalies in the cities/countries. To this end, we consider some factors such as symptoms of the disease, mortality rates, temporal patterns, peak and recovery of the disease, timing of outbreaks, numbers of people infected and recovery ability of the individual cities.

KEYWORDS: visual analytics, investigative analysis, intelligence analysis, sense-making, analysis process

INDEX TERMS: I.3.8 [Computer Graphics]: Applications-Visual Analytics, I.6.9 [Visualization]: information visualization, H.5.2 [Information Systems]: Information Interfaces and Presentation.

1 INTRODUCTION

Version 5.1 of Tableau was used for solving Mini Challenge 2. Tableau, data visualization software [1], was a project in Stanford University in 1997 by Chris Stolte [2] and Professor Pat Hanrahan [3] which was released in 2003. With Tableau, analysts can connect to the large amounts of data and auto-generates visualizations including reports, tables, charts and graphs through a drag-and-drop interface.

Tableau uses a database visualization language called VizQL (Visual Query Language) [4]. It combines a structured query language for databases with a descriptive language for rendering graphics. It specifies views, as well as describes tables, charts, graphs, etc. Due to the switching from one visual representation to another (e.g. from a list view to a cross-tab to a chart), visualizations can be easily customized and controlled.

VizQL is used by Show Me [5], a set of user interface commands and defaults, to extend automatic presentation to tables of views. User experience, included the automatic selection of mark types, a command to add a single field to a view, and a pair of commands to build views for multiple fields has been described [5].

SIAT, Simon Fraser University, 102 Avenue, Surrey, Canada
Send correspondence to Minoo.E.J.
Minoo.E.J.: E-mail: mea18@sfu.ca
Mona.E.J.: E-mail: mea16@sfu.ca
C.D.S.: E-mail: shaw@sfu.ca

Additionally, design study of graphical history tools for Tableau has been described in [6]. They explained a design space analysis of both architectural and interface issues, including identifying design decisions and associated trade-offs. They then elaborated on the tools for recording and visualizing interaction histories, supporting data analysis, and outlining mechanisms for presenting, managing, and exporting histories.

1.1 ANALYSIS PROCESS

Microsoft Excel was used for combining the measured attributes of all eleven cities/countries. Again the excel file was inserted to Tableau for further analysing and comparing cities/countries.

We loaded two files of each eleven cities/countries to Tableau. The provided files were in the form of .csv with varied sizes from 90K records to 7 million records. Using Tableau was efficient regarding the process of analyzing the data as well as the work needed to be done to enter the data, in particular for the huge datasets like Karachi and Aleppo.

In the section 2 we describe the key features that we used during our analysis.

1.2 CHALLENGES

Generally we faced with a few challenges during the analysis process. As we were only modestly familiar with Tableau, we needed explore its facilities, and occasionally repeated analyses as we explored the challenge dataset. Once we developed an analytical approach for one or two of the countries, we were able to repeat this approach for the remaining countries in the dataset.

2 TABLEAU FEATURES

Tableau's features made the analysis of data straightforward and easy. Followings are the main features we used.

2.1 JOIN

We used the join feature where the attribute ID was the primary key in the two files of each city/country. For each countries/cities we left joined the two files in order to have all the attributes of patients and deaths in one table. Sometimes it took up to 4 minutes to join large files in Tableau.

2.2 FILTERING

Tableau's filtering feature was essential for our analysis. We were able to set different conditions on the fields to be filtered. We set several conditions on the Symptoms, Number of Records, Date of Death, etc. For instance we filtered the data to visualize only the patients in the outbreak with 92 symptoms which had a high frequency to find the symptoms of the outbreak. We can easily include or exclude groups (e.g. patients) by setting different conditions.

2.3 Calculated Field

Another useful feature was defining a new calculated field. For example we add a function named "DATEDIFF" to calculate the number of days between two dates. DATEDIFF was used to measure the number of days between hospital admittance of dead people and their date of death, i.e. the hospitalization time for the dead people. Therefore we realized the hospitalization time for patients died in the outbreak was exactly 8 days. As a result, we deduced that if a patient in the outbreak could survive more than 8 days he/she then would recover.

2.4 Visualization Type

There were a lot of options to select the type of visualization; we mostly used bar chart and scatter plot in our analysis [See Figure 1 and 2]. We also used pie chart and line chart. Moreover, we easily could deal with different types of data such as Date, Number, String, etc. Additionally each field could have different formats for its values e.g., the Date fields in MC2 had mixed format of 1/11234, august 2 2001, 1 august 1999. Other features that we used were colour coding, size and sorting the bar charts.

System also provides a summary for each visualization which consists of some information of the visualized data such as count (number of records), sum, average, maximum and minimum.

From bar charts we could see the exact dates clearly. For example, Patients in the outbreak had hospital admittance date starting April 16th until June 29th with a peak in from May 13th through 21st. Similarly, those who died in the outbreak had hospital admittance date starting April 16th-22nd until Jun 18th-26th with a peak during May 14th-20th.

We could easily sort bar charts ascending or descending. For example we realized 'NOSE BLEEDS' had the highest number of people infected which was almost double the rest of 91 symptoms by sorting the graph ascending.

Color coding and size coding were other useful features that we can add to visualization to have a more meaningful visualized data. For instance we color coded the data by the month of admittance date, death date, gender, etc. So we realized that age and gender were not significant factor as they were normally distributed in all the countries/cities.

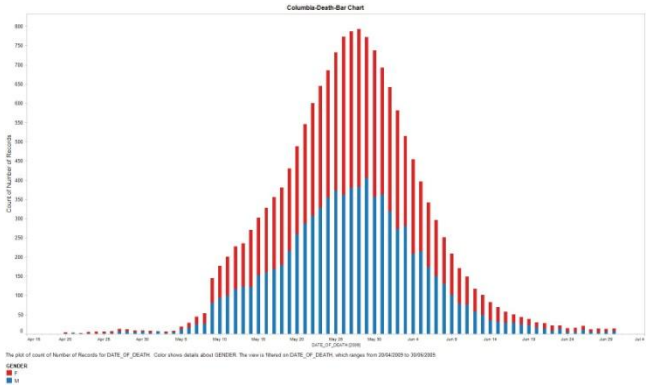

Figure 1. Bar chart with colour coding. Columbia dataset; vertical axis is the number of records and horizontal axis is the date of death. The colour represents the gender (red for female and blue for male).

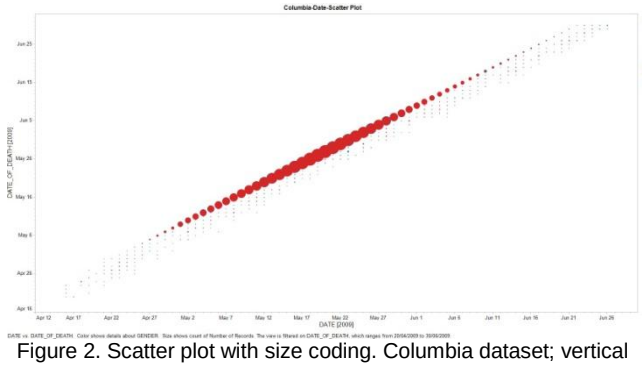

Figure 2. Scatter plot with size coding. Columbia dataset; vertical axis is the date of death and horizontal axis is the date of admittance. The size of bubbles represents the number of records.

2.5 Dashboard

Dashboard was another useful feature in our analysis. It enables us to insert different worksheets (visualizations) into one panel and compare those to each other. We could select part of the data in one visualization and saw how the selected data appear in other visualizations. Therefore for each country we had a dashboard of its different visualizations.

3 Conclusion and Future Work

Our main approach with Tableau was to address the question of the timing and location of disease outbreak. These questions were driven largely by questions arising from the Grand Challenge, in which we viewed disease source and timing as central issues. We thus focused our efforts on temporal issues. Our goal was to link disease progress to the travel of arms dealers from Dubai in MC1.

References

[1] Tableau software http://www.tableausoftware.com/

[2] http://graphics.stanford.edu/~cstolte/

[3] http://graphics.stanford.edu/~hanrahan/

[4] Pat Hanrahan, VizQL: a language for query, analysis and visualization, Proceedings of the 2006 ACM SIGMOD international conference on Management of data, June 27-29, 2006, Chicago, IL, USA [available online at http://portal.acm.org/citation.cfm?id=1142473.1142560]

[5] Jock Mackinlay , Pat Hanrahan , Chris Stolte, Show Me: Automatic Presentation for Visual Analysis, IEEE Transactions on Visualization and Computer Graphics, v.13 n.6, p.1137-1144, November 2007 [available online at http://portal.acm.org/citation.cfm?id=1313046.1313140]

[6] Jeffrey Heer , Jock Mackinlay , Chris Stolte , Maneesh Agrawala, Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation, IEEE Transactions on Visualization and Computer Graphics, v.14 n.6, p.1189-1196, November 2008 [available online at http://portal.acm.org/citation.cfm?id=1477066.1477414&coll=GUIDE&dl=GUIDE&CFID=98802757&CFTOKEN=95154988]