CZSaw, IMAS & Tableau: Collaboration among Teams

VAST 2010 Grand Challenge Award: Excellent Student Team Analysis

Dustin Dunsmuir, Mahshid Z. Baraghoush, Victor Chen, Minoo Erfani Joorabchi, Mona Erfani Joorabchi, Saba Alimadadi, Eric Lee, John Dill, Cheryl Qian, Chris D. Shaw, Robert Woodbury

School of Interactive Arts and Technology, Simon Fraser University

ABSTRACT

The VAST 2010 Challenge consisted of three separate datasets which we investigated with three student teams using three different tools in order to solve each Mini Challenge (MC1-3). The teams met to share findings, request supporting evidence from the other mini challenges, and raise questions for other teams to investigate further. We used CZSaw [1] to investigate the MC1 arms dealer reports by organizing an overview before drilling down to investigate each country's activities. We used Tableau for MC2 to summarize the spread of the Drafa virus within each country and compare the times at which it occurred. We used the IMAS [2] genomics tool for MC3 to discover the origin and initial spread of the virus.

KEYWORDS: Visual analytics, investigative analysis, information visualization, analysis process, biology and genetics

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; J.3 [Computer Applications]: Biology and Genetics

1 INTRODUCTION

For this Grand Challenge we initially worked in three separate groups which subsequently collaborated to share results. In the team meetings, each team described their findings. Following this, we brainstormed possible connections between scenarios and developed two competing hypotheses. Further investigation with the tools allowed us to narrow this to a single supported scenario with evidence connected between mini challenges by common geography and paired time frames. Using IMAS, we determined the country of origin of the virus. Using Tableau we developed a timeline showing the spread of the disease to other countries. The MC1 data in CZSaw described the activities of people from each of these countries and allowed us to connect the other two challenges with a scenario involving arms dealers spreading the disease from the origin country to the others. Our hypothesis was strengthened by matching dates from arms dealer meetings discovered with CZSaw to disease outbreaks analyzed in Tableau. In the following sections we provide an overview of our use of these tools, focusing on their novel features and the discoveries we made about them during the analysis.

2 CZSAW – MINI CHALLENGE 1

CZSaw is a visual analytics tool that focuses on capturing and supporting the analysis process. CZSaw contains data views for visualization and manipulation of entities, entity collections, and relations. CZSaw also creates an editable, re-playable, and re-useable script of the analysis process to help analysts understand, explore, reference and reuse this process. The script contains the model of the analysis and provides the foundation for a visual *History View* of the process and a *Dependency Graph*.

CZSaw visualizes entity collections; thus we asked our colleagues in the SFU Natural Language Lab to run entity extraction algorithms on the original dataset. Unfortunately this challenge highlighted the fact that such algorithms are never perfect and refinement is required on the extracted entity set. To help with this, we built into CZSaw a capability to alter the set of entities by extracting new ones and removing unnecessary ones as well as merging and linking entities. For example, when a person's name was spelled differently and thus tagged as two separate entities, we could easily merge these together. We also discovered connections between phone numbers and their owners that could be encoded as a person entity with a phone alias.

CZSaw's *Dependency Graph* maintains dependencies among all data results generated during the analysis. Each data result, a node in the dependency graph, is dependent upon the root data or an existing data result. Altering the root data or content of a result causes propagation in the *Dependency Graph* resulting in an automatic update of the data views. This quick update of the data views allowed us to smoothly integrate data exploration with refinement of the data.

The data views within CZSaw provided us the flexibility we needed to organize the dataset into subsets for each country and then drill down into these subsets. CZSaw's *Semantic Zoom View (SZV)* is a document cluster overview with a focus+context continuous zoom. Initially the *SZV* displays all documents in the dataset as small rectangles. A layout algorithm places documents containing many of the same entities close to each other. From this overview, a semantic zoom can be applied to any subset of documents to see in detail their set of entities and then their full text. During this zooming, the SHriMP algorithm [3] is used to move the other documents (the context) outward while the focus documents are expanded. The *SZV* also allows brushing and linking with entities and grouping documents. We were able to quickly divide the data into groups by country and then each could be analyzed by a single team member. These groupings were stored within CZSaw's script which allowed team members to not only re-generate the same groups by running the script on their individual machines but also to step through the process of analysis that was taken to create them. To investigate the social network of arms dealers in each country and across countries, groups of documents could be dragged from the *SZV* and dropped into a *Hybrid View*.

CZSaw's *Hybrid View* displays relations between document and entities in a node-link graph. It can display nodes (entities) in different forms and layouts. By populating a *Hybrid View* with entities related to one of the subsets of documents we could see the network of people and places in a country. The force directed

E-mail: { dtd, mzeinaly, yvchen, mea18, mea16, salimada, ela10, dill }@sfu.ca, qianz@purdue.edu, shaw@sfu.ca, rw@sfu.ca

layout in this view clusters entities and documents by pulling closely related nodes together by their edges. The edges between clusters informed us of the connections between disparate activities within each region and led us towards understanding the larger scenario.

3 TABLEAU – MINI CHALLENGE 2

Using Tableau for MC2 was effective for importing and analyzing the data, in particular for huge datasets (for Karachi and Aleppo). The features of Tableau made the analysis of data straightforward and easy. Below we describe the key features used.

We used the join feature where the attribute ID was the primary key in the two files of each city/country. We left joined the two files in order to save all the attributes of patients and deaths. However, it took up to 4 minutes to join large files in Tableau.

Tableau's filtering capability was essential for our task. We were able to set different conditions on the fields to be filtered. For instance we set several conditions on the Symptoms, Number of Records, Date of Death, etc. We could easily include or exclude groups by setting different conditions.

Another useful feature was the definition of new calculated fields. For example we added a function named "Datediff" to the attributes. It was used to measure the number of days between hospital admittance of the deceased and their date of death, i.e. the hospitalization time for the deceased.

There were many options for selecting the type of visualization; we predominantly used bar charts and scatter plots in our analysis. Moreover, Tableau easily handled 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 20/8 2010, August 2010, or 20 August 2010. Other features that we used were colour coding, size and sorting the bar charts.

4 IMAS – MINI CHALLENGE 3

IMAS (The Interactive Multi-genomic Analysis System), a visual analytics system for the discovery of knowledge in genomic information, was used to analyze the genomics data to determine the evolution of the Drafa virus. IMAS was initially developed by Shaw and his students in 2007. IMAS is available at http://imas.sourceforge.net under the GPL 3 license. IMAS enables the user to load various FASTA format files. One or more sequences can then be selected to work with at a time.

This tool visualizes the output of common bioinformatics tools such as BLAST and ClustalW in a unified framework. In this challenge, BLAST was used for pair-wise nucleotide sequence alignment. Pair-wise alignment visualizes the character-by-character similarities and color highlights the differences between sequences.

IMAS provides us with a horizontal zooming of the sequences. This interaction assists in the discovering of patterns in the whole sequence area by controlling the level of detail. Those patterns emphasize non-conserved regions at a glance. We can then zoom in to a region of interest for more detail.

4.1 Finding Pair-wise Relationships

When determining the original country of all the virus strains, we assumed that the native sequence which displays the most similarities to each of the current outbreak sequences was the ancestor of all the mutant strains. We defined the similarity between two strains as the number of different bases; those sequences that have the least number of base substitutions are the most similar.

Fig. 2 shows one BLAST run result for sequence strain 118 against all the countries. The light purple areas show the similar regions and the green rectangles highlight the differences, which indicate that at least one substitution takes places at that area. (We ignored green color gradients).

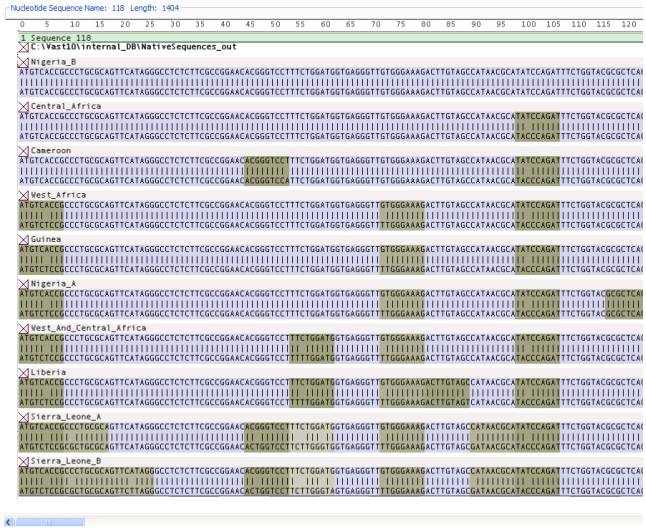

Figure 2. The results of the pair-wise alignment of sequence 118 against each of the countries.

By looking at the image it is clear that the first sequence, Nigeria-B, has the least number of differences with the strain 118 in comparison with others.

5 CONCLUSION

The tools we used in our analysis were specific to the data we were given for each mini challenge. CZSaw excelled at handling the textual data's multiple threads of activity through entity refinement, semantic zooming, clustering, grouping and node-link diagrams.

Tableau, was easy and effective for MC2 and we suggest using it for the next epidemic outbreak**.**

With IMAS, we found the virus origin country in a few minutes. The clear visualization of the result of pair-wise alignment of each strain against all the countries helped us to observe that the least number of mutation areas was for Nigeria_B. Although IMAS was successfully used in solving the tasks and determining the correct answers, we believe the interaction aspects of the tool could be further improved.

These three tools performed efficiently at this VAST challenge, and we now better understand them for future improvements.

REFERENCES

[1] N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw and R. Woodbury. "Capturing and Supporting the Analysis Process", Proceedings of IEEE Visual Analytics Science & Technology, (Atlantic City, NJ, Oct 11-16, 2009), pp. 131-138, 2009.

[2] C. Shaw, G. Dasch, and M. Eremeeva. "IMAS: The Interactive Multigenomic Analysis System", Proceedings of IEEE Visual Analytics Science & Technology, (Sacramento, CA, Oct 30-Nov 1, 2007), pp. 59-66. 2007.

[3] M-A. D. Storey, and H. Muller, "Graph Layout Adjustment Strategies", Proceedings of the Symposium on Graph Drawing, (Berkeley, CA, Sept 18-20, 1996), pp. 487-499. 1996.