VAST 2006 Contest - First Place, Student Category
Visualizing Relationships in a Diverse Data Collection

Kanupriya Singhal[*]
Georgia Institute of Technology

Summer Adams[†]
Georgia Institute of Technology

Abstract

This paper is a summary of the contest entry submitted to the VAST contest 2006. The primary task of the VAST contest is to determine what, if any, illegal activities are happening in the fictitious town of Alderwood, Washington. The paper summarizes the data analysis performed on the synthetic data set, introduces the visualization tool, and describes the strengths and weaknesses of our approach. The tool we developed facilitates exploration of the diverse data by providing the user with multiple visualizations of the data that are tightly integrated. More specifically, we integrate a Treemap, node-like graph, timeline, and map as well as provide a table based view into the complete data set. The user is able to filter the data as they desire which results in all relevant visualizations being updated.

Our submission to the contest began as a class project. After the class ended, we continued work on the application for approximately one additional month. During this time, we enhanced some of the functionality and fixed defects. We also completed further detailed analysis of the VAST contest data.

CR Categories: H.5.2 [User Interfaces]: Interaction styles; I.3.6 [Computer Graphics]: Methodology and Techniques-Interaction techniques; J.m: Miscellaneous

Keywords: Text, graphs, maps, social networks, user interaction, VAST contest

1 Introduction

The goal of the 2006 VAST contest is to determine if inappropriate activities are taking place in Alderwood, Washington based on a synthetic dataset. The dataset consists of 1200 news articles, photographs, maps and some textual materials. Given the quantity and nature of the data involved, we decided to devise a set of views that help analyze such datasets. Our visualization tool consists of four distinct views, a Treemap [1] view, a classic node-link graph view, a timeline and a map view. The rest of the paper discusses the data analysis task, the design of the visualization tool and its interactions. The strengths and weaknesses of the tool are discussed followed by a conclusion.

2 Data Analysis

One of the most challenging aspects of the contest is the data analysis. Familiarization with the data required considerable effort. Evaluation of the data set led us to define five separate entity types; Person, Organization, Location, Event, and Activity. These entities encompass most of the information and actors within the dataset. Scripts were written to extract entities and their occurrence counts from the data set. Stemming using the Porter stemmer [2], stop wording, and text frequency-inverse document frequency (TF-IDF) [3] were employed to remove extraneous information. All entities extracted were ultimately converted into formats appropriate for each visualization we created. An XML file was generated for the initial Treemap view. Subsequent XML files were generated at run-time based upon filter criteria for detailed views within the Treemap and the graph view. Matrix files were produced to capture the entity-entity relationships and entity- news stories relationships.

3 System Design

The overall system design is based on the principles of brushing and linking. The visualization provides multiple co-ordinates views. The Treemap view on the top left and the filter panel on the right, provide the overview of the entire dataset. The filter panels are populated based on the existing entities categorized into each entity set.

Based on selections in the filter panel, the map visualization at the bottom left of the application, displays location information. A textual view of the documents and articles is provided in the top center panel of the application window. A timeline visualization shows the temporal relationships among entities, articles and their publish date.

The central view in the application is a node-link graph from the prefuse package. Based on selections in the filter panel, the node-link graph view is updated with a network diagram. Related entities are connected by edges and the colors of the nodes identify the entity type. If two selected entities are unrelated, the node-link graph indicates this by showing each entity without an edge between them. Hovering over a node reveals the entity name.

In addition to these visualizations, an entity list is provided. This facilitates searching and sorting.

4 System Interaction

The system is driven by the detailed data on the right of the screen. By single clicking on one of these entities, the related entities are highlighted in the other tables. Since the list of entities may be long, we created a side panel that corresponds in location to the highlighted rows of a table. Clicking on one of the green

[*] e-mail: ksinghal@cc.gatech.edu

[†] e-mail: summer@cc.gatech.edu

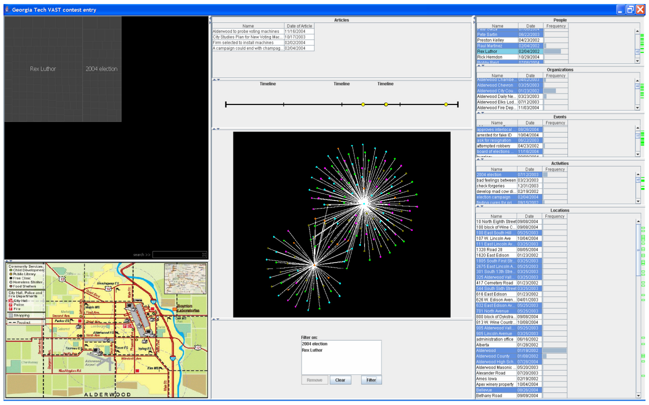
bars forces the entity table to jump to the location that was selected. By double clicking on an entity list item, the entity list item will appear in the query table, which is at the bottom center of the screen. Choosing to filter will cause all the visualizations to be updated with respect to the items that are in the query table.

The tree map shows the entities that have been filtered upon and the relationships that entity has with other items. The sizes of the nodes represent the occurrence frequency giving us an idea of the relative importance of an entity in the pool of filtered entities. The nodes in the Treemap are highlighted when moused over. The user is able to search for items in the Treemap by using a search tool specific to the Treemap panel.

As for the map visualization, the filter only shows relationships that are locations on the map. If the filter does not produce any locations, then the map visualization will not be updated and will stay static.

The social network visualization is also affected by the filter. The social visualization shows the entities that each item is related to and how they are related to each other.

For the articles, when the filter button is selected the articles that are related to all of the filtered entity items are displayed in the table. This information provides the user with direct access to the articles and the photos that the relationships are gathered from. For instance, if a user wants to understand the relationship that mayor has with the city council, then he or she could filter on these two entities and only read the articles that contain information on both the mayor and city council.



Articles

Name	Date of Article
Alderwood to probe voting machines	11/16/2004
City Studies Plan for New Voting Mac...	10/17/2003
Firm selected to install machines	02/02/2004
A campaign could end with champag...	02/04/2004

Filter on:
2004 election
Rex Luthor

Remove | Clear | Filter

Figure 1 – Visualizing a relationship in the application. This figure shows the filtered data for data entities Rex Luthor and the 2004 election.

In addition to the table of articles, the temporal timeline provides information about the articles. In fact, the temporal timeline provides a list of the filtered articles with respect to time, such that users should be able to quickly gather when the articles were written and an estimate of the amount of articles written without heavy calculations. The user is able to mouse over the timeline to see the article title and the date the article is written.

As an example, take the case of selecting the individual Rex Luthor. The user would first single-click his name in the entity table list and gets a feel for the number and type of connections to Rex Luthor. The user may then double-click the name and have it added to the query list in the bottom middle pane. The user may then choose to filter on Rex Luthor which would update the views. The treemap would display Rex Luthor as the primary topic and each entity he is linked to as a node in the treemap. The map view would highlight a map location if one on the map corresponded to one he is linked to. The timeline and articles would update to show all articles he is mentioned in. And finally, the network graph would display Rex in the center with links to all other entities.

To take it one step further, the user may then choose to add another item to the filter list and re-filter based on both (or more as the user selects) entities. Again, all views update appropriately with the articles treated as an AND of the query entities.

5 STRENGTHS AND WEAKNESSES

We believe our application has a number of strengths. The application provides the users with an overview of the entire dataset. Starting from a complete picture, the filtering mechanism facilitates exploratory data analysis. The filter panels coupled with multiple coordinated views allow investigators to gain insight into the data at different levels of details. The complementary visualizations are much more powerful together than any single visualization. The visualization relies heavily on the relationships among the various entities. Having multiple views highlights these relationships and the filtering can lead to valuable insights. And finally, our visualization application is practical. Visualization techniques prevalent in the community since the past several years have been incorporated into this tool.

One weakness with the application is the data extraction. Much work was done to extract information from all data sources without human intervention; however some sources still required hand manipulation. Another limitation of the system is that at present we have not provided any provision to visualize the entire entity network at once. But we can easily add that if need be.

6 CONCLUSION

We combine some existing visualization techniques in a new and interesting way allowing investigators to explore entity relationships within datasets. The extracted relationships along with the coordinated multiple views serve as a concrete visual analytic technique to assist investigators in analyzing communities such as Alderwood.

7 ACKNOWLEDGMENTS

We would like to thank Susan Gov and Sheena Lewis for their valuable contributions to the initial class project. We would also like to thank Dr. John Stasko for his feedback and insights.

REFERENCES

[1] Bruls, M., Huizing, K. and Wijk, J. J. v. (2000). *Squarified Treemaps*. Proc. of VisSym '00 (May 2931, Amsterdam, The Netherlands), SpringerWien NewYork, 33-42.

[2] Porter. 1980."An algorithm for suffix stripping." Program, Vol. 14, no. 3, pp 130-137. http://www.tartarus.org/_martin/PorterStemmer

[3] Salton, G. and Buckley, C. "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, Vol. 24, No. 5, pages 513-523, 1988.