Story Explorer: A Visual Analysis Tool for Heterogeneous Text Data

Roy G. Biv*
Starbucks Research

Ed Grimley†
Grimley Widgets, Inc.

Martha Stewart‡
Martha Stewart Enterprises
Microsoft Research

ABSTRACT

In this paper, we propose Story Explorer , a visual analytic system for text data from multiple sources. With various visualizations, our system can help analysts identify conflicts and correlations in large volume of text data, and detect patterns of group of people. Thus analysts can discover the development of events and find the suspicious people in the events.

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 INTRODUCTION

Exploring heterogeneous text data from different sources can be complicated and with error pruning if not well dealing with hidden relationship between entities or possible data conflicts between materials. In the MC1 of VAST Challenge 2014, to reveal potential relationship between POK and GAStech, we need to extract important people from different sources of files, e.g. employee resumes, email headers, research reports, and a huge volume of news with conflicts. In our design of the analysis tool Story Explorer , we integrate different kinds of source files into a timeline-based view, providing a quick overview for users to choose important file to focus on for efficiency.

In this paper, we will first introduce our design consideration of visual analysis tools for heterogeneous text data and then introduce how we use these tools to analyze MC1 data and to solve these problems.

2 DESIGN CHALLENGES

Data provided in MC1 mainly includes 845 news articles, 35 resumes, employee records, 1171 email headers and some other reports on both POK and Kronos. As the data volume is so large, we need an efficient overview to put important data together. Main challenges that we face in our design of Story Explorer are listed below:

Data correlation and data conflicts In MC1, data are greatly overlapping and thus there exists correlation and conflicts. For example, both resume and employee records present GAStech employees' information, and reports on POK and Kronos cover some information in news articles. Therefore we have challenges to present data visualization: On the one hand, our tool must enable users to extract information from multiple sources as they do contain the relationship which we need. On the other hand, we need to give obvious hints for data conflicts and provide details to help the users resolve these conflicts.

*e-mail: roy.g.biv@aol.com
†e-mail:ed.grimley@aol.com
‡e-mail:martha.stewart@marthastewart.com

Manipulating data at high level MC1 data in news articles and email headers cannot be easily presented due to their large volume. Existing tools like Jigsaw and Google Fusion are great to deal with text visualization, however, they don't provide exploration in a higher level and thus users won't have a quick start to focus on data which they are interested in. So challenge exists in providing the users with high level manipulation of data to understand data distribution or the trend of development before reading detailed data with Jigsaw or Google Fusion.

3 VISUALIZATION TOOLS

In this section, we will introduce our tool design and how we deal with challenges mentioned above.

Our tool Story Explorer includes Resume-Reader for GASTech employee information, News-Timeline for news articles and Email-Reader for identifying mailing communities from email headers. We will explain our design of these three views in detailed below.

3.1 Resume Reader

VAST Challenge 2014 focuses on rescuing missing people, thus employees of GAStech play an significant role in the whole event. Resume Reader integrates employee records with their resume, and allow users to identify suspicious people according to conflicts presented in the view and identify potential suspicious groups by reading their experience timeline. The goal we want to achieve with this tool is to expose text conflicts and to identify potential communities in GASTech.

Exposing Conflicts We think that conflicts between resume and employee records may expose potential forged resume to help identify suspicious people. By integrating temporal information from two sources into experience timeline, Resume Reader provides a clear view for users to identify text conflicts. Experiences are presented as small squares and important dates are highlighted in the timeline, thus users can identify conflicts and then refer to detailed description to check it.

Identify Potential Relationship Another thing which we consider important in the design of Resume Reader is to enable users to identify potential relationship based on common working or education experience, as common experience in the same place for a long time may lead to the formation of a community. When these people are all in GAStech, it's likely that they will form a group of their own. Dragging and filtering function are designed for this consideration.

Figure 1: Identify conflicts and check it using detailed description

3.2 News timeline

News timeline is designed to provide a quick view for the users to grab the development of an event and then focus on certain points to analyze.

As the volume of news data is large, simply displaying news in a timeline may be imbalance and will result in a lot of overlapping in a short period of time. So semantic zooming is used in News Timeline, which makes it possible to provide a compact view in the timeline when the time range is long and provide news position precisely when the range is short.

Aside rooming function, News Timeline provides visual set operation on timelines. Every time a user searches a keyword, a new timeline will be generated and displayed on the screen. As timeline operations are allowed for users, a user can do set operations between them to refine the result he/she finds. Draggable dialogs are also provided to refine the data: a user can pick up important news he/she finds and arranges them in the timeline by dragging to prepare for further investigation.

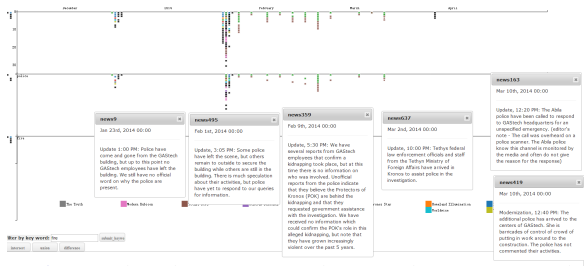

Figure 2: Operational NewsTimeline

3.3 Email Reader

{**TODO: Is that complete?**}

Design And Overview In MC1 we have email headers from two weeks of internal GAStech company email, we can get a social network from this data and then discover communities. For MC1, we need to reveal the connections between GAStech employees and to find suspectable clues including the subject of emails. If there is an email containing words related to POK, then we can try to find the connections between the sender of the email's community and POK. Therefore we implemented a visual analytic tool for email headers based on D3.js. The tool is easy to use and effective to discover communities.

Layout And User Interactions Our tool's layout containing four components, filters, email sending and receiving timeline, email headers view, and community view. After selecting an employee through the filter, his/her email records will be depicted on the timeline, the contents of email headers will be put in the email headers view and some communities including him/her will be showed in the community view.

Users can interact with the layout both directly and indirectly, including selecting employees, filtering by keywords and limiting the size of groups. When a user consider a keyword or a person to be suspicious, he/she can easily see people who send the suspicious emails or are close to the suspicious person.

4 Data Exploration

In this section, we introduce how we use Story Explorer along with analysis tools like Jigsaw, Gephi to analysis MC1 data.

As our tool only provides an overview of the whole data, we still need to use Jigsaw to assist the analysis process. And general analysis includes following steps: 1. Read report on POK and GAStech to identify important people involved in the conflicts. 2.

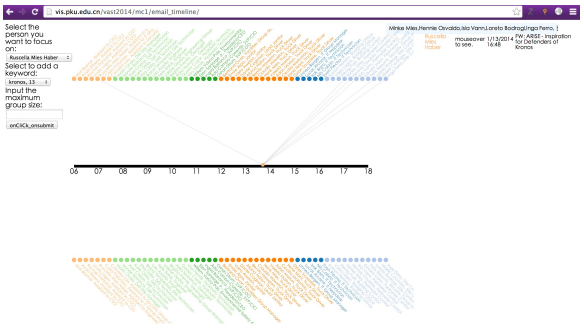

Figure 3: Email Reader

With important entities identified, using News Timeline to refine keywords. 3. Cooperated with Jigsaw to read news articles and find relationship between POK and GAStech. 4. Discover important communities and identify suspicious people through Resume Reader and Email Reader. 5. Go on for another round of exploration. 6. Present our find with the help of Gephi.

And with rounds of explaration, we can finally get enough information to answer questions.

5 Conclusion

Story Explorer provides the user with the ability to grab important events from huge volume of news ariticles and the ability to analyze conflicting data of employee resume. And with the help of Story Explorer, we successfully find out a group of GAStech people who are suspicious to the kidnap event. Our methods emphasize analyzing various sources of text data, and in the future we will focus on integrating the visualizing tools in Story Explorerto enable users to switch views conveniently when they find some people suspicious through one tool, and want to confirm their suspicions by data from another source. {**TODO: Add more conclusion.**}

Acknowledgements

We have special thanks to Miss Dong Liu for her dedicated exploration with raw materials to inspire our design on visualization tools.

References

[1] G. Grinstein, D. Keim, and M. Ward. Information visualization, visual data mining, and its application to drug design. IEEE Visualization 2002 Course #1 Notes, October 2002.

[2] G. Kindlmann. Semi-automatic generation of transfer functions for direct volume rendering. Master's thesis, Cornell University, 1999.

[3] Kitware, Inc. *The Visualization Toolkit User's Guide*, January 2003.

[4] M. Levoy. *Display of Surfaces from Volume Data*. PhD thesis, University of North Carolina at Chapel Hill, 1989.

[5] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Computer Graphics (Proceedings of SIGGRAPH 87)*, volume 21, pages 163–169, July 1987.

[6] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, June 1995.

[7] G. M. Nielson and B. Hamann. The asymptotic decider: Removing the ambiguity in marching cubes. In *Visualization '91*, pages 83–91, 1991.

[8] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, second edition, 2004.