Story Explorer: A Visual Analysis Tool for Heterogeneous Text Data

Chenglong Wang Zhengjie Miao Siming Chen Zipeng Liu Zuchao Wang
Zhenhuang Wang Xiaoru Yuan[*]

Peking University

Abstract

In this poster, 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.

1 Introduction

Exploring heterogeneous text data from different sources can be complicated and misleading if not well dealing with hidden relationship between entities or possible data conflicts between materials. To reveal potential relationship between POK and GAStech in MC1, we need to investigate important information using files from different sources, e.g. employee resumes, email headers, research reports, and a huge volume of news. To solve these challenges, we integrate different kinds of source files into our timeline-based analysis tool Story Explorer. It serves as a quick overview tool for original data, allowing users to focus only on important files for efficiency.

In this paper, we will first introduce design challenges of visual analysis tools for heterogeneous text data and how we design Story Explorerto solve them. Then we will introduce how we use Story Explorer to solve MC1 problems.

2 Design Challenges

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

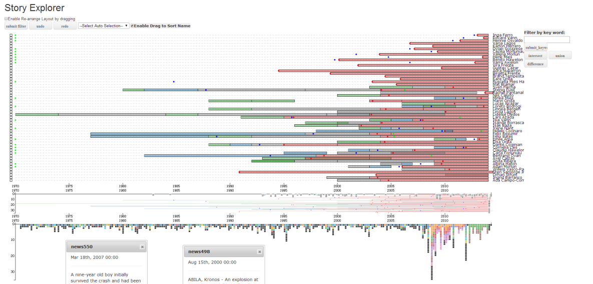
Data correlation and data conflicts In MC1, data are greatly overlapping and thus there exists correlation and conflicts. For example, both resume and employee records present GAStech employees' information, and reports on POK and Kronos cover some information in news articles. Therefore we have challenges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

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.



news550

Mar 18th, 2007 00:00

A nine-year old boy initially survived the crash and had been flown to Abila Children's

news498

Aug 15th, 2000 00:00

ABILA, Kronos - An explosion at a nearby gas drilling well pad

Figure 1: An overview of Story Explorer

3.1 Resume Reader

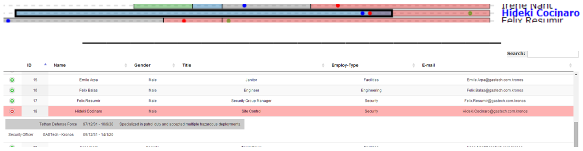
VAST Challenge 2014 focuses on rescuing missing people, thus employees of GASTech play an significant role in the whole event. Resume Reader integrates employee records with their resume, and allow users to identify suspicious people according to conflicts presented in the view and identify potential 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.

[*] e-mail: {chenglongwang, miaozhengjie, csm, zipeng.liu, zuchao.wang, wangzhenhuang, xiaoru.yuan}@pku.edu.cn

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE



	ID	Name	Gender	Title	Employ-Type	E-mail
⊕	15	Emile Arpa	Male	Janitor	Facilities	Emile.Arpa@gastech.com.kronos
⊕	16	Felix Balas	Male	Engineer	Engineering	Felix.Balas@gastech.com.kronos
⊕	17	Felix Resumir	Male	Security Group Manager	Security	Felix.Resumir@gastech.com.kronos
⊖	18	Hideki Cocinaro	Male	Site Control	Security	Hideki.Cocinaro@gastech.com.kronos

Tethan Defense Force 97/12/31 - 10/9/30 Specialized in patrol duty and accepted multiple hazardous deployments.

Security Officer GASTech - Kronos 09/12/31 - 14/1/20

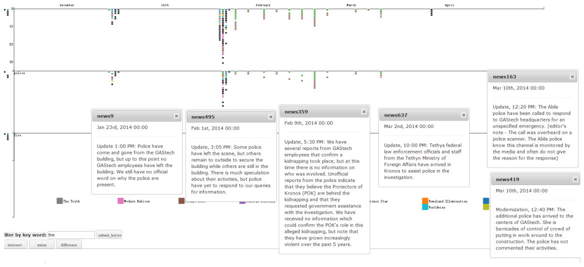
Figure 2: Identify conflicts and check it using detailed description

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.



news9
Jan 23rd, 2014 00:00
Update 1:00 PM: Police have come and gone from the GAStech building, but up to this point no GAStech employees have left the building. We still have no official word on why the police are present.

news495
Feb 1st, 2014 00:00
Update, 3:05 PM: Some police have left the scene, but others remain to outside to secure the building while others are still in the building. There is much speculation about their activities, but police have yet to respond to our queries for information.

news359
Feb 9th, 2014 00:00
Update, 5:30 PM: We have several reports from GAStech employees that confirm a kidnapping took place, but at this time there is no information on who was involved. Unofficial reports from the police indicate that they believe the Protectors of Kronos (POK) are behind the kidnapping and that they requested government assistance with the investigation. We have received no information which could confirm the POK's role in this alleged kidnapping, but note that they have grown increasingly violent over the past 5 years.

news637
Mar 2nd, 2014 00:00
Update, 10:00 PM: Tethys federal law enforcement officials and staff from the Tethyn Ministry of Foreign Affairs have arrived in Kronos to assist police in the investigation.

news163
Mar 10th, 2014 00:00
Update, 12:20 PM: The Abila police have been called to respond to GAStech headquarters for an unspecified emergency. (editor's note - The call was overheard on a police scanner. The Abila police know this channel is monitored by the media and often do not give the reason for the response)

news419
Mar 10th, 2014 00:00
Modernization, 12:40 PM: The additional police has arrived to the centers of GAStech. She is barricades of control of crowd of putting in work around to the construction. The police has not commented their activites.

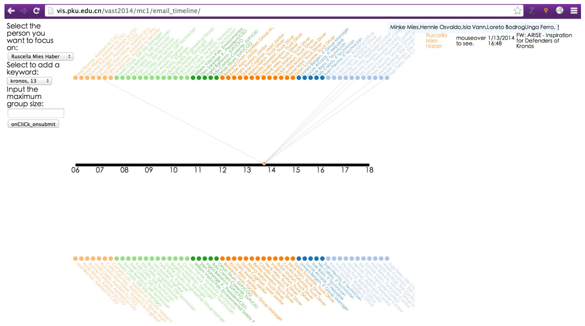
Figure 3: Using operational News-Timeline to pick up important events

3.3 Email Reader

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 suspicious 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, e-mail 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 e-mail 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.


vis.pku.edu.cn/vast2014/mc1/email_timeline/
Select the person you want to focus on:
Ruscella Mies Haber
Select to add a keyword:
kronos, 13
Input the maximum group size:
onCliCk_onsubmit
06
07
08
09
10
11
12
13
14
15
16
17
18

Minke Mies,Hennie Osvaldo,Isia Vann,Loreto Bodrogi,Inga Ferro,

Ruscella Mies Haber	mouseover to see.	1/13/2014 16:48	FW: ARISE - Inspiration for Defenders of Kronos

Figure 4: Using Email Reader to identify the employee relationship

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. With important entities identified, using News Timeline to pick up important events associate with them.

3. Cooperated with Jigsaw and Fusion 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. Display the result by putting important news in the timeline and draw the relationship network with Gephi.

With rounds of exploration, we are able to identify suspicious people and pick up events related to the employee missing events. Results are presented in forms of timeline events and relationship net work to provide for further investigation in Grand Challenge.

5 Conclusion

Story Explorer provides user the ability to extract important events from huge volume of news articles and the ability to analyze conflicting data of employee resume.

With the help of Story Explorer, we successfully identify a group of GAStech people who are suspicious to get involved in the kidnap event. Our visual analysis tools emphasize on analyzing text data from different sources and provide an overview of the whole dataset.

Visualization tools to display results we found are also necessary and that can be part of future work.

Acknowledgements

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

References

[1] J. Stasko1, C, Grg1 and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization Information Visualization (2008) 7, 118 – 132

[2] David Jonker, William Wright, David Schroh, Pascale Proulx, Brian Cort. Information Triage with TRIST 2005 Intelligence Analysis Conference, May 2005, Washington, DC.