ClueMiner: A real-time multi-dimentional visualization system

Vast Challenge 2014: Honorable Mention for Good Support for Streaming and Forensic Analysis

Siqi Yang*
Visual Analytics Lab, Tianjin University

Xinyi Jiang†
Visual Analytics Lab, Tianjin University

Jiawan Zhang‡
Visual Analytics Lab, Tianjin University

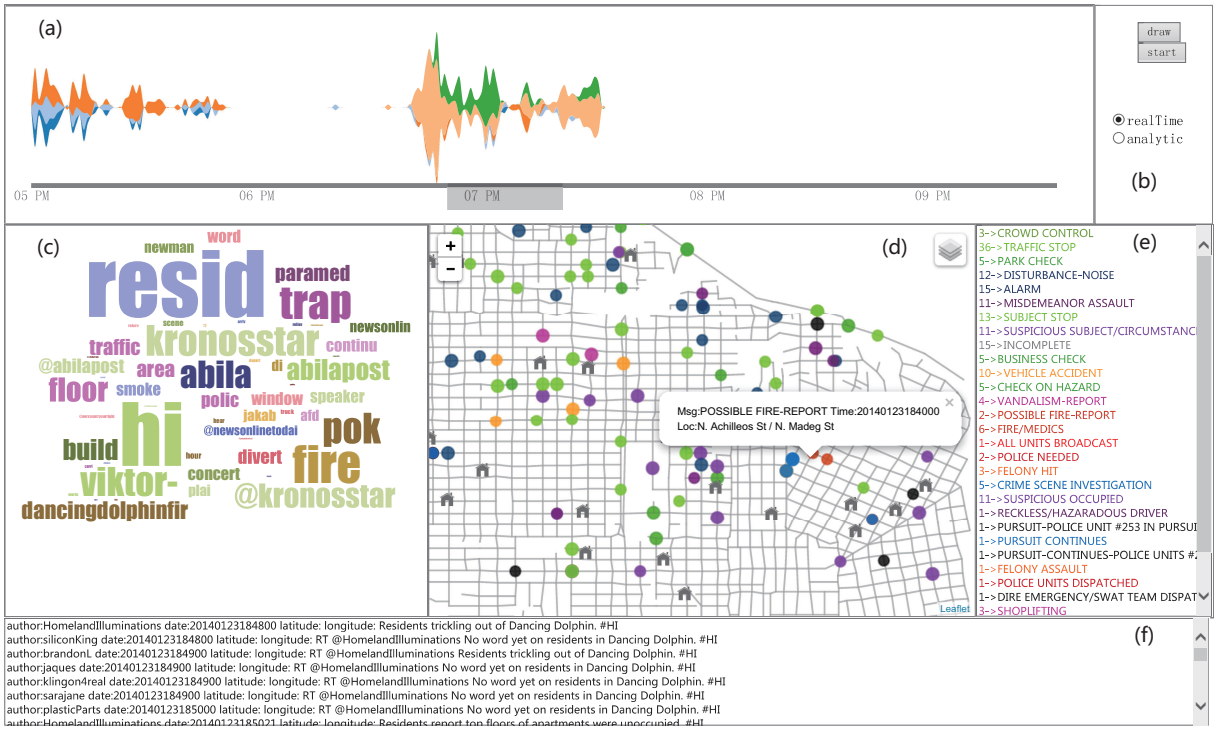


Figure 1: Interface of our ClueMiner. (a)Stream View; (b)Control View; (c)Word Cloud; (d)Map (e)Statistics View; (f)Metadata View.

Abstract

In this paper we briefly describe the tool submitted to the mini-challenge 3 of VAST Challenge 2014. The tool enables users to explore text stream and acquire geographic information in real time without prior knowledge of the data. In addition, we design other tools to assist the author analysis of twitter. The combination of tools also provides good interaction operation allow users to analyze the twitter data in real time.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces; H.1.2 [User/Machine Systems]: Visual Analytics;

1 Introduction

The provide data for VAST 2014 mini-challenge 3 is text stream include twitter data and call-center data which contain author, message information, time and geographic location. To make sense of the data, its important to make analyst see these different dimensions of the data. Also, analyzing such amount of stream text, the analyst should allow finding out and investigating specific events with the tool. To meet these requirements, we developed the tools to visualize and analyze data from three main aspects: author visualization and analyzes geo-temporal visualization, and time-varying twitter message visualization. In the next sections, we describe how the tools were used in the resolution of the VAST Challenge.

*e-mail: yangsiqi@tju.edu.cn
†e-mail: akanea@163.com
‡e-mail: jwzhang@tju.edu.cn

2 Our Method

First we will introduce our analytic approach. As we care more about the reliability and efficiency of the real-time analytic tool, analyses were conducted on the data in segment 1 and segment 2. Then the analysis of streaming data based on the result of pre-analysis is in real-time.

2.1 Analysis of segment 1 and 2

The study of segment 1 and 2 are from three aspects.

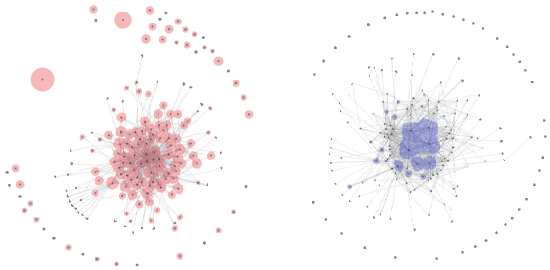
Data ontology The data contains two different types: twitter data and call-center data. The content of twitter data has lots of irrelevant information. Taking into consideration of the twitter is written by someone, we tried to filter the information by identity of

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

the author. As for the call center data, we categorize the message column into five level according to the risk level.

Author Identity The re-tweet relationship regarded as a proper measure of author identity in a social network. During identity analysis, we found two typical categories which are illustrated in Fig.2.

Event Taking the correlation of twitter messages of different authors into consideration, we could identify the similarity of the authors and interests of them may the potential events in progress.



(a) (b)

Figure 2: Re-tweet relationship.(a)Nodes has larger size but no links are junk accounts; (b)Larger nodes in center are official accounts or important participants.

The whole analysis process is described below. For the twitter data, we first filter out the junk messages by using author identity analysis. We use re-tweet relationship to measure the author identity. Taking the number of twitter of each author into account, we generate a force-based node-link map show the re-tweet relationship of authors.

In Fig.2(a), the link shows the re-tweet relationship of two authors, the size of each node refers to the number of twitter of specify author. As it shown, there are some big nodes with no link with other authors which means no one pay attention to what they said, so the messages from these authors were not important that we treated them as junk messages. In Fig.2(b), the authors in the center of the graph who take more space are the most re-tweeted user. They turned to be the official account or a people who directly attend in an event.

After all the data received, we provide a view to present the similarity between authors. We collect all the twitter messages of a specify author into a single document to find out the similarity of these documents. A simple and computationally cheaper method is to calculate the similarity using word counts using term frequency-inverse document frequency (TF/IDF). After get the similarity matrix, we use multidimensional scaling(MDS)[2] to draw a scatter plot, each document is a node represents an author. MDS makes similar documents are placed close to each other as Fig.3 shown.

2.2 Analysis of stream data

We pay more attention to the analysis of the stream data and provide the analysis and presentation on the real time. The analysis consists of two dimensions, spatial and temporal information. We provide three main views in our analytic tool: map view, word cloud and stream view.

Stream data in the spatial distribution All call-enter data and some twitter data contain the geospatial information. All data contain the time information. We resolve the address information and mark the places in the map real-time. In Fig.1(d), we marked all the call center data. Colors represent the different level of a matter from green to red. Green represents the lightest alert. On the

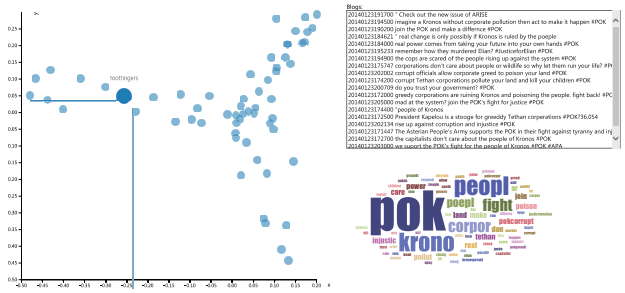


Figure 3: Scatter plot view represents the similarity of authors.

contrary, red represents the most serious scene. Analysts can select any levels of the alert, and see the detail information of each mark.

Stream data in a timeline Text analysis is the most important part of our system. We receive the stream data from the browser, and preprocess the data real-time. The time-slice is one minute in our system which means we deal with the text message streams together in each one minute, with the visualization update. We use general method to do the text preprocess which include set stop words, stem to merge the words which have different forms, update vocabulary, and record word count. In visualization, we simply choose Word Cloud shown in Fig.1(c) and Themeriver[1] shown in Fig.1(a) to display the events. The displayed words in Word Cloud are the most increased words in a specific time range which is defined by using a sliding window. A stream river is used for presenting the change of word frequency over time. Words in the stream view also can be selected by users' interaction.

2.3 Interaction

As a real-time analytic tool, it needs some time for processing the stream data, so a system delay is set, which is one minute in our experiment. Our analytic tool has two modes which are real-time mode and analytic mode. When receiving the stream data, the system is in real-time mode. As it finishes receiving, it change to analytic mode where we can do much deeper study on the data. However, the real-time analysis is exceedingly important to the whole problem, interaction in real-time mode as follows.

The system display real-time data without any interactions. When you brush the stream view, you select a specific time range and information in map view and words in word cloud refresh. Focus on a specific category of call center data which shown in statistics view in Fig.1(e) to update the map. For more detail information on the map, click the mark on the map or zoom in and out. By clicking the words in word cloud, update the stream view. But when a time range is assigned, by clicking the words, the metadata view shows the original twitter message and call center data.

3 Conclusion

With our tools, the interconnected three main component: author visualization and analyzes geo-temporal visualization, and time-varying twitter message visualization, worked together to help analyst explore the stream data from different aspects in real time. The tool is suitable for stream data analysis and crime forensic matching the demands of VAST 2014 mini-challenge 3.

References

[1] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[2] W. S. Torgerson. Theory and methods of scaling. 1958.