VAST 2014: Summary on Mini Challenge 3 Work

Minjing Mao*
ASTRI

ABSTRACT

VAST 2014 Mini-Challenge 3 requires to identify the events for further investigation on the disappearance of GAStech employees from the streaming data. To tackle this mini-challenge, we proposed a visual analytics tool. In this paper, we summarize the design and implementation of our tool. Besides, we illustrate how we use the tool to indicate the events.

Keywords: VAST 2014 Challenge, Mini-Challenge 3, visual analytics, natural language processing (NLP).

Index Terms: [Information systems applications]: Data Mining—Data Stream mining; [Human-centered computing]: Visualization—Visualization application domains—Visual analytics

1 INTRODUCTION

VAST 2014 Challenge focuses on a fiction incident in the country of Kronos. Tethys-based gas company GAStech has made huge profits via operating a natural gas production site in Kronos. However, GAStech has not paid attention to the environmental pollution caused by its project. In the celebration of its new-found fortune on January 20th 2014, several employees of GAStech went missing. Protectors of Kronos (POK), one local environment protection organization was suspected in the disappearance. To find the missing employees, three mini-challenges as well as the grand challenge are set up to analyze the data from various aspects. Specifically, Mini-Challenge 3 provides the micro-blog messages and call center records from 1700 to 2130 Abila time on January 23, i.e. three days after the incident. The focus is to identify the events of interest for further investigation through analyzing the streaming data [1].

To analyze the streaming data efficiently, we propose a tool that supports the real-time processing. Besides, it can also support off-line processing on the data from batch files. For the development, we use Java and PostgreSQL [2] with a JavaScript library, i.e., Data-Driven Documents (D3) [3].

The rest of the paper is organized as follows. In Section 2, the design and implementation of our proposed tool is introduced. The results and analysis procedure will be presented in Section 3. Finally in Section 4, we summarize our findings.

2 SOFTWARE

In order to unearth an event and its timeline in Mini-Challenge 3, two major issues need to be considered in the design and implementation of our tool.

One issue is how to abstract the key components of an event from the messages: who, how, where and when. First, we apply the Standford NLP [4] to preprocess every message to eliminate the stop-words, stem and tag the remained words. Second, we count the co-occurrence of any two words in every minute to aggregate the words' relationship. We then measure the importance of each word by counting the number of messages that contain it. This is because the more messages contain the word, the more attraction it receives. After aggregating the data, we put the result into the PostgreSQL database for future visual analytics. Finally, we calculate the Levenshtein distance between any two messages, and remove the same and similar messages. This is to clean up the result for display.

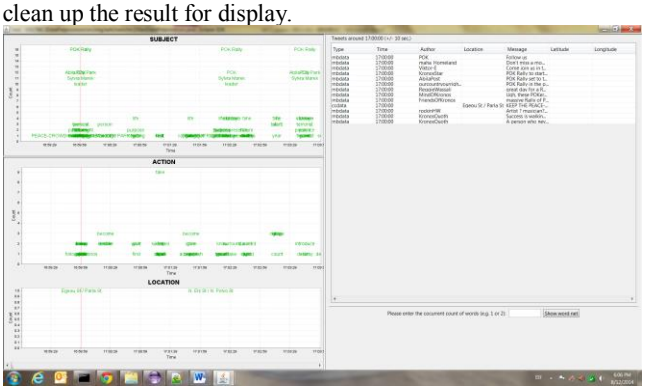

Figure 1: Main Graphic User Interface.

The other issue is how to properly display the key components summarized from the previous processing. To solve this, we mainly use three charts: SUBJECT, ACTION and LOCATION (as shown in Figure 1). Specifically, we put time on x-axis and word count on y-axis; the word itself appears on the chart. Notably, all three charts are synchronized. To facilitate the better analytics, we set a marker (the red line in each chart) to indicate the time that the user is currently viewing. Through all the word charts, we can easily tell the high frequent subjects, actions andlocations, respectively. By dragging the scrollbar below the charts, we can view all the data that are already processed. The charts are updated periodically when newly aggregated data are available. In this way, we can see the timeline and progress of events. For those words which are overlapped, we can either zoom in the chart, or check the details of messages in the right panel. The right panel lists the messages around the current time being viewed.

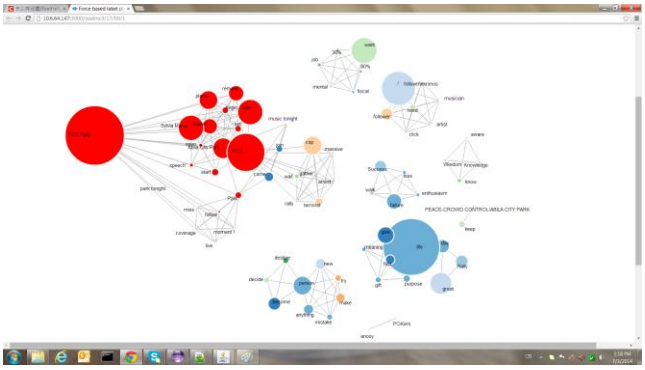

Figure 2: D3 Bubble Chart.

* minjingmao@astri.org

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

To better visualize the events, we also use D3 bubble chart to show the relationship of words and their importance (as shown in Figure 2). After we enter the word co-occurrence in the main interface, the bubble chart will be created based on the data of current minute bin in the database. In the D3 bubble chart, the size of each bubble shows the word count/importance, and the link in between stands for the co-occurrence of the words. When we are checking one word's relationship, we can easily click the corresponding bubble, and then all the related bubbles will be highlighted in red.

3 Result And Analysis

Before we present the result and analysis, we briefly introduce how we use our proposed tool to find an event. Take the first event of POK Rally for example. As shown in Figure 3, the chart of subject gives us some top frequent subjects: "POK Rally", "Sylvia Marek", "Abila City Park", "Prof. Stefano", "Lucio Jacab", "Victor-E" and "Dr. Newman". We then try to find the verbs corresponding to the above subjects in the chart of action. Together with using the word net in D3 bubble chart, we can indicate the related actions are: "open", "begin", "introduce", "speak" and "play". From the chart of location, we can see the place of "Egeou St / Parla St". With studying the call centre record that contains the place (as listed on the right panel), we can confirm it is the event location.

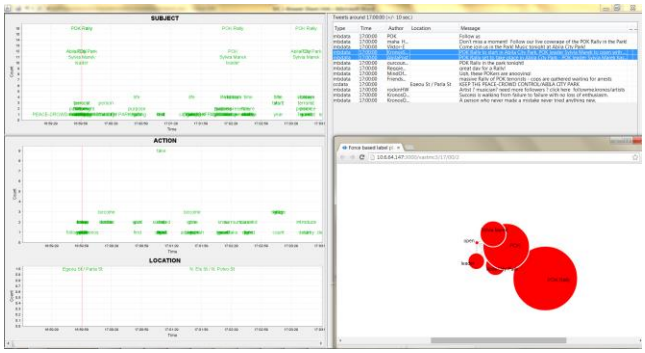

Figure 3: POK Rally.

In this way, we can abstract an important event from the raw data: POK Rally. It started at 17:00:00 and ended around 19:05:54. The progress of the event can be described as follows: the leader of POK, i.e. Sylvia Marek, gave an opening speak to start the rally in Abila City Park; a special guest Dr. Newman was invited to give a talk with Prof. Stefano and Lucio Jacab; the band Victor-E was invited to play music.

Similarly, we abstracted three other important events: the fire at the Dancing Dolphin Apartment, the vehicle accident between a black van and a bicycle, and the gunfire and subsequent events after the black van was stuck at the Gelato Galore. Due to the limit of length, we do not repeat the progress and timeline of all the four events in this paper, which have already been described in the submitted answer sheet.

We now illustrate why we consider the fire at the Dancing Dolphin Apartment most likely to provide additional clues to the investigation of the GASTech disappearances. The fire started at around 18:40:00, and did not end until 21:30:00 (the end of streaming data). Its progress can be summarized below: the smoke was found from first two floors; police cleared the area and rescued the residents inside; a firefighter injured and was sent to the hospital; this fire was reported to be under control; the top floors have collapsed into the first floor later, and the fire was confirmed suddenly escalated; fire officials have begun control measures on the surrounding buildings; the fire was still not completely controlled, and an explosion was reported at 21:30:00.

Figure 4: Fire at the Dancing Dolphin Apartment.

As shown in the right panel in Figure 4, we find a witness "dangermice" reporting the whole progress of fire. This people's microblog data include the longitude and latitude. With such information, we can check the fire location on the tourist map via ArcGIS [5] (as plotted in Figure 5). From Figure 5, we can see that the location of vehicle accident and the fire is quite closed. In addition, the vehicle accident happened just a few minutes after the fire. Then we can make a reasonable assumption that the people who hit the bicycle may be in a hurry to escape the crime scene of the fire at Dancing Dolphin Apartment. Since the suspect in the vehicle accident has been caught, the further investigation may emphasize more on the fire.

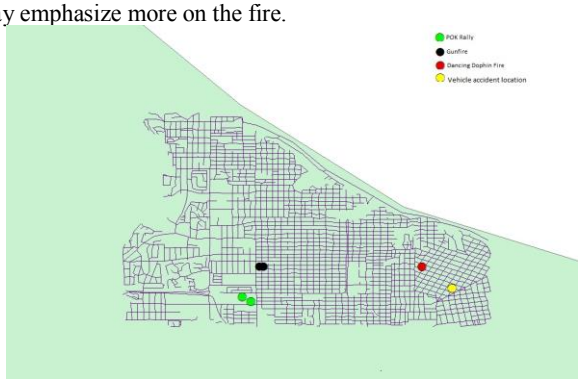

Figure 5: Location of Events.

On the other hand, we find another witness "Onlythetruth", who expressed his or her unsatisfactory to GasTech using the words of "burn the foreign capitalist pig". It implies that the Dancing Dolphin Apartment is close related with the company GasTech. Based on the above analysis, we believe that the fire at the Dancing Dolphin Apartment is most likely to provide important clue for the future investigation.

4 Conclusion

We proposed a tool to solve the VAST Mini-Challenge 3. It can support both real-time and off-line processing. Our proposed tool efficiently abstracted the key components of events and properly displayed them. The result shows that our proposed tool can effectively help solve the problems in this mini-challenge.

References

[1] http://vacommunity.org/VAST+Challenge+2014
[2] http://www.postgresql.org/
[3] http://d3js.org/
[4] http://www-nlp.stanford.edu/
[5] http://www.esri.com/