Visual Analytics of Text Streams Through Multiple Dynamic Frequency Matrices

Nicolas Médoc*
LIPADE, Université Paris Descartes
Centre de Recherche Public Gabriel Lippmann

Mickaël Stefas†
Centre de Recherche Public Gabriel Lippmann

Mohammad Ghoniem‡
Centre de Recherche Public Gabriel Lippmann

Mohamed Nadif§
LIPADE, Université Paris Descartes

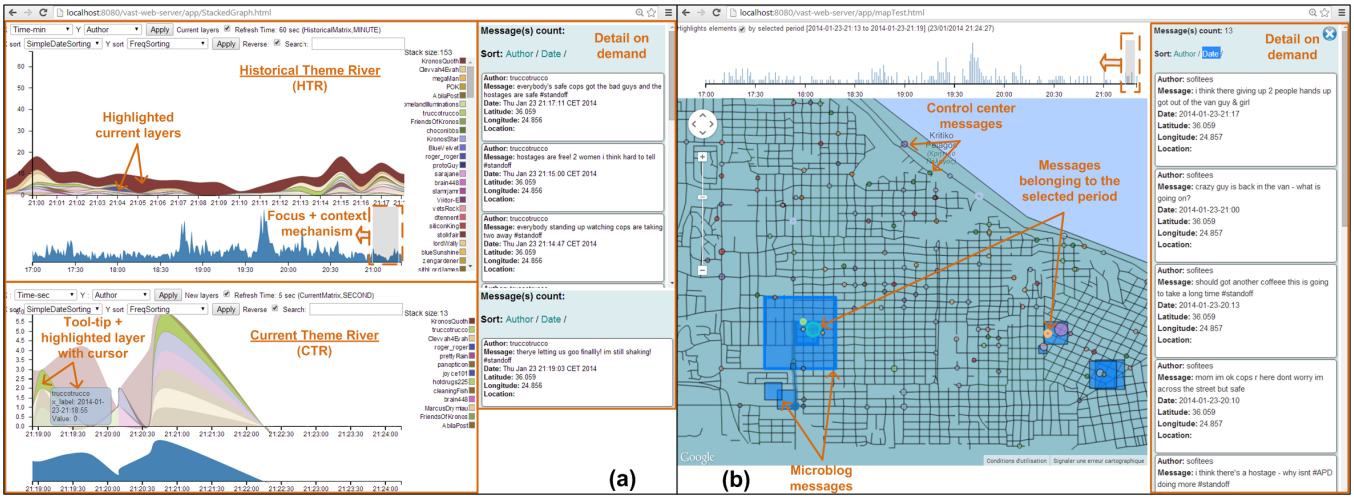

Figure 1: **(a)** Two dynamic Theme Rivers (or Stacked Graphs) show the evolution of different aspects of streaming messages from a microblog. **(b)** A map of Abila City is displayed through the GoogleMap API. The *Control Center* (CC) messages are represented by circles and the *Microblog* (MB) messages, at their reported coordinates, are represented by squares.

Abstract

We propose a Visual Analytics tool that supports situation awareness and exploration tasks for text streams. To reach this goal, we design our own data model to encode streaming text in multiple dynamic frequency matrices, handling multiple aspects of data. Our visualizations are composed first of two dynamic Theme Rivers. They allow real time exploration of all the aspects extracted from texts stored in both, short-term and long-term buffers. In addition, we visualize the geographical location of messages on a map. We use these visualizations, enhanced by efficient user interaction mechanisms, to answer the questions of the third mini-challenge of 2014 VAST Challenge. As a result we identify several challenging issues that we will investigate in future work.

Keywords: Visual analytics, text stream, vector space model.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interface;

*e-mail: medoc@lippmann.lu
†e-mail:stefas@lippmann.lu
‡e-mail:ghoniem@lippmann.lu
§e-mail:mohamed.nadif@mi.parisdescartes.fr

1 Introduction

The third mini challenge (MC3) of 2014 VAST Challenge focused on the field of Visual Analytics of text streams. The provided data consist of messages from a microblog similar to Twitter and alert messages located in a city, supplied by the fire and the police departments. The purpose of the challenge was to discover important events, their location and the principal users involved in a fictional terrorism scenario. Our visual analytics tool allows the analyst to reach the MC3 objectives by carrying out situation awareness and exploration tasks. In *Section 2* we present related work that has inspired our approach, our data models, and our visualizations. We describe briefly in *Section 3* our software architecture and our data model, then our visualization design in *Section 4*. Finally we discuss in *Section 5* the issues that we encountered during this work.

2 Related Work

Three main previous contributions have inspired us to develop our tool. Firstly, we based our design on the following tasks proposed by Rohrdantz et al [4]. The first task consists of *monitoring the current situation* by a short-term analysis of data. An *exploration* task consists in navigating through all dimensions of data using interactive visualization, allowing the user to fully understand multiple aspects of data. *Event tracking* as well as *change and trend detection* tasks allow the user to follow real-time evolution of data. The *historical retrieval* task allows the user to analyze the historical development of a selected subset of data. Finally, the *temporal context* task analyzes the temporal evolution of ranks regarding different aspects of data [3] and relates them to the current situation.

Secondly, one of the most used data models for text corpora is the

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

Vector Space Model (VSM), also known as *Bag Of Words*. Based on the *VSM*, Turney and Pantel [5] identified three models to handle three kinds of matrices: the *Term-Document* matrices, *Word-Context* matrices and *Pair-Pattern* matrices. We propose a more general framework described in *Section 3*.

Thirdly, Stacked Graphs [1] are visualizations derived from the Theme Rivers [2]. We enhanced it by interaction mechanisms to support the tasks previously mentioned.

3 Architecture and models

Event-driven architecture for web application. We have adopted a web-oriented architecture using Apache Tomcat for the back-end, and D3.js and AngularJS for the front-end. We use Esper, a Java component for Complex Event Processing, to handle streaming data in multiple buffers with flexible expiration strategies. We have defined a long-term buffer for history analysis and a short-term buffer to monitor the current situation.

Text processing. We have chosen to exploit in priority structured contents available in microblog messages such as metadata, hashtags and mentioned users. We used Twitter's Text Processing Library for the extraction of the Twitter's structure. Furthermore we used Named Entity Recognition algorithms from Stanford CoreNLP API to extract meaningful elements from messages such as persons, locations and organizations.

Model for multiple dynamic frequency matrices. Although many java libraries exist for modeling matrices or vector space models, we choose to develop our own in order to ensure the following requirements. The model must compute multiple dynamic frequency matrices, at least one for each desired aspects. The state of matrices must represent the messages belonging to associated long-term or short-term buffers. To reach this purpose we defined two kind of spaces: the *Object Space* and the *Context Space*. In *Object Spaces* the entry values correspond to any feature available as metadata or extracted from messages. In *Context Spaces*, the entry values correspond to features of data in which items of *Object Space* can appear many times. In this way, *Object-Context* matrices can be built by combining instances of both *Spaces* and by counting the *Objects* in each *Context*. For example, we can use hashtags as *Objects* and time in minutes as *Context*. Several instances of *Spaces* can be defined and several matrices can be associated to each instance (e.g. Author-TimeInMinute, Author-TimeInSecond or Hashtag-TimeInMinute). Each *Space* entries and each matrix weight are maintained as messages stream in. When a message in a buffer expires, we adopt different removal strategies depending on how the matrices are affected and how we want to propagate this effect to the spaces.

4 Visualization design

Dynamic Theme Rivers. In *Figure 1(a)*, the Historical Them River (HTR) is dedicated to historical analysis and shows how text streams have evolved in the long-term buffer. The Current Theme River (CTR) provides situation awareness in real time and gives insight into text streams in the short-term buffer. In the configuration area, the analyst can combine one instance of *Context space*, mapped to the X axis, with one instance of *Object space*, represented by the stacked layers on the Y axis. As a result, the evolution of matrix weights are visible through the thickness of layers over time. Their colors are mapped to the labels showed in the scrollable legend. Therefore, the analyst can observe the temporal development globally for one selected aspect (i.e. authors, hashtags or named entities), achieving *temporal context* task. At regular intervals, the new states of the selected matrix are propagated to the views. Through these dynamic Theme Rivers the analyst can hence *monitor the current situation* and *detect changes*. The HTR layers are automatically highlighted when the same item exists in the CTR. This supports the *historical retrieval* task. In contrast, the CTR layers are highlighted when they correspond to new items. This achieves *trend detection* task. When the cursor passes over the Theme Rivers or over the legend, the related layer is highlighted and a tool-tip provides label information. This is useful for the *exploration* task. In addition, while the Theme Rivers give an overview of the temporal patterns, we added a *Focus and Context* mechanism to allow the analyst to zoom and explore in detail the Theme River in a selected period that can be dragged along the time-line. Finally, the legend can be filtered and the original messages are displayed on demand by clicking on a layer or on a legend item.

Location of events in a map. In *Figure 1(b)*, the low number of events located in the city allow all of them to be displayed, giving an historical overview. The bar chart shows their distribution on a time-line covering the long-term buffer. Each message can be highlighted on the map with a thick border to achieve two objectives. The first helps to *monitor the current situation* by observing the items belonging to the configurable sliding window preceding the current time. The second approach consists in selecting a period in the bar chart. By dragging it on the time line, the items are successively highlighted when they belong to the period. This helps to carry out *exploration* and *change and trends detection* tasks.

5 Discussion and future work

During our work, the first issue we encountered concerned the CTR visualization where the analyst monitors the current situation only for one aspect at a time. Nevertheless, it would be useful to have multiple aspects monitored simultaneously in order to have a full understanding of the situation. The second issue concerns our *historical retrieval* solution based on data stored in memory. However, for long periods or massive streaming, the available memory can be insufficient. This means that more sophisticated storage solution and information retrieval techniques must be considered so that the history remains efficiently accessible for the analyst. In future work, we will apply other text processing techniques to uncover hidden structure of streaming text. We will investigate such techniques while keeping in mind the trad-off between the need to have efficient user interaction and the need to have accurate results, to achieve situation awareness in real time. Moreover, many visualizations have been proposed for analysis of complex structures in static text corpora (i.e. topics hierarchies or semantic networks). We would like to investigate their suitability and ways to adapt them for streaming text.

6 Conclusion

Our Visual Analytics tool, with the ability to handle dynamically multiple aspects of text streams, allowed us to answer MC3 questions. Beyond the participation in the VAST Challenge, this work has paved us the way for experimenting novel visualizations and text processing techniques for streaming texts that we aim to investigate, in order to tackle the issues identified in this paper.

References

[1] L. Byron and M. Wattenberg. Stacked graphs-geometry & aesthetics. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1245–1252, 2008.

[2] S. Havre. ThemeRiver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[3] M. Krstaji, M. Najm-Araghi, F. Mansmann, and D. A. Keim. Story tracker: Incremental visual text analytics of news story development. *Information Visualization*, 12(3-4):308–323, July 2013.

[4] C. Rohrdantz, D. Oelke, M. Krstajic, and F. Fischer. Real-time visualization of streaming text data: Tasks and challenges. In *IEEE VisWeek*, volume 201, 2011.

[5] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141188, Jan. 2010.