Eurographics Conference on Visualization (EuroVis) (2014)
N. Elmqvist, M. Hlawitschka, and J. Kennedy (Editors)

The Stor-e-Motion Visualization for Topic Evolution Tracking in Social Media Streams

Andreas Weiler[1], Michael Grossniklaus[1], Franz Wanner[2], and Marc H. Scholl[1]

firstname.lastname@uni-konstanz.de
[1]Database and Information Systems Group, [2]Data Analysis and Visualization Group, University of Konstanz, Germany

Abstract

Nowadays, there are plenty of sources generating massive amounts of text streams in a continuous way. For example, the increasing popularity and the active use of social networks results in voluminous and fast-flowing data streams containing a large amount of user-generated data about almost any topic around the world. However, the observation and tracking of the ongoing evolution of topics in these unevenly distributed text streams is a challenging task for analysts, news reporters, or other users. This paper presents "Stor-e-Motion", a real-time visualization to track and explore the ongoing evolution of topics' frequency (i.e., importance), sentiment (i.e., emotion), and context (i.e., story) in user-defined topics over continuous flowing text streams.

1. Introduction and Motivation

The high volume and distribution speed of tweets makes it difficult for users to follow the evolution of topics within the continuous data flow. It is a big challenge to discriminate between normal behavior of the social sensor or unusual and abnormal behavior, which usually is an indicator for an interesting event in the area. However, the amount of useful information in the generated data increases as well. In this paper, we present an application for visually tracing and monitoring the importance, emotion, and story of user-defined topics in the live and continuous data stream of Twitter. Our work presents a compact visualization for time series event data, which supports users in identifying interesting data points in the large volume of tweets. Additionally it is possible to overview whole sets of topics and to compare the evolution of different topics with each other over time. Furthermore, the application automatically displays interesting terms in a tag cloud fashion over time. In this work, we focus on the visualization of text streams from the social microblogging platform *Twitter*, for which we present a use case applied on real-world data streams collected from the public timeline of Twitter.

2. System Design

It is a big challenge to discriminate between normal behavior of the topic evolution or unusual and abnormal behavior, which usually is an indicator for an interesting event in the context of a topic. Therefore, the visualization is tailored to support the characteristic of fast distribution and spreading of information of social media services. In the following, we describe the three design goals - visualizing the evolution of the *Importance*, *Emotion*, and *Story* of a topic.

Importance: The importance of the topic in the time window is visualized by using the size of the shape. Since the length of a time window is pre-defined and unchangeable, we use this value as the width of the rectangle. For the height of the rectangle we calculate for each shape the value against a pre-defined *max_size* with the values *n* (total search hits inside the window) and *m* (total items inside the channel and window) by using the *Inverse Document Frequency* [SJ88] in the following formula:

$$size = \left(\log\left(\frac{n}{1}\right) - \log\left(\frac{n}{m}\right)\right) * \left(\frac{max\_size}{\log\left(\frac{n}{1}\right)}\right).$$

Emotion: The emotion of a topic is visualized by using the coloring of the shapes. The filling color of the shape signifies the average sentiment (red = negative, blue = positive, yellow = neutral) of the text in the data window. The value of the sentiment for a text segment is calculated by using an external library from Thelwall *et al.* [TBP*10], which analyzes the text of each tweet and returns sentiment values between -5 (extremely negative) and 5 (extremely positive). To

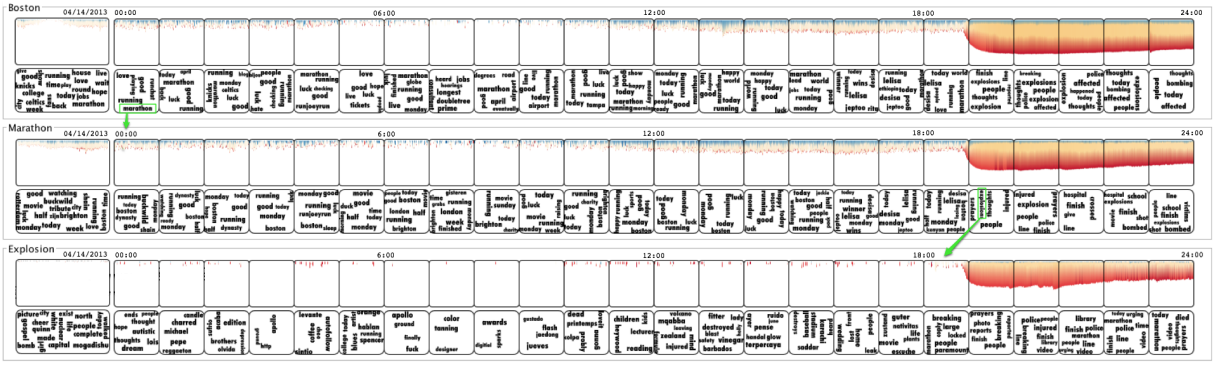

Figure 1: *Topic visualization (top to bottom) for "Boston", "Explosion", and "Marathon" from April 15th 2013. Support of real-time visualization every minute, automatic roll-up to a daily compressed view, and drill-down into the compressed views.*

visualize the ratio of positive and negative sentiment in the data windows, we use a linear gradient coloring with a color map from [Bre12] and the overall percentage of positive or negative tweets in the data window.

Story: We visualize the story by using a continuously changing tag cloud (minute wise) for the real-time visualization, which becomes static after one hour of processing to reflect the tags of the corresponding hour. The tag cloud is created by summarizing the most frequent terms, which are co-occuring with the term of the topic definition. We filter out terms included in a standard English stopword list and the topic definition term. Hereby we ensure that only terms are displayed, which have a certain influence to the story.

Interaction: A topic can be defined by using a single keyword. The ongoing evolution of the story around a topic, which eventually deals with important subtopics around the main topic, mostly triggers the interest in new topics, which did just occur and are missing in the original topic definition. Therefore, analysts are able to just click on single terms in the tag clouds of any topic and a new topic analysis starts automatically. Figure 1 shows that the visualization automatically rolls-up after a 24-hours of processing to an aggregated visualization, which reflects the whole day in a more compressed way. However, users are able to manually drill-down into the day and see a more detailed view of the topic.

3. Use Case

In this use case, we describe the observation of the city *Boston* on the day of the 15th April 2013. By using the Twitter Streaming API [Twi] with the so-called "Gardenhose" access level, we receive 10% of the public live stream. The dataset for that day contains a total of 53 million tweets, which is an average of about 2.2 million per hour and about 37,000 tweets per minute. Figure 1 shows the visualization of three topics for the whole day. The first topic *Boston* is defined by using the keyword "boston". By observing the ongoing evolution of *Boston* we can see that the frequency decreases after a couple of hours. By that time, it was night in Boston and the people sending tweets about *Boston* are mostly asleep. By looking at the co-occurring terms of the story of the topic *Boston*, we can derive that there is an ongoing event called *Marathon*. Since we are also interested in that event, we add a new topic definition by clicking on the term *Marathon* in the tag cloud. The context in both topics is mostly about "running" and "marathon". We can also see that when the "marathon" is finished the winners "jeptoo" and "desisa" are mentioned. The most interesting pattern appears a few hours after the first runners finished the marathon. The negative emotion of both topic increases and drifts into extremely negative. The overall importance of the topics increases significantly and therefore reflects the happening of an interesting event. At this point in time an *explosion* took place at the finishing line of the marathon course. By clicking at the term *explosion* a new topic is defined and added to the visualization. We can also see that "line" appears in the tag cloud. Shortly after new terms like "prayers", "police", "injured", and "obama" appear, which reflects the current situation very well.

4. Conclusions

The use case shows that the visualization supports users in keeping the overview in following and tracking topics over time and also guides the users to interesting points or periods in time. Furthermore, the visualization supports the user in the situational awareness around a topic and in serendipitous findings. These findings can be easily used to create new topics. Future work includes an evaluation and a user study of the visualization, to find out if the visualization really supports users in identifying events in topics and are able to understand the happenings in their defined topics. An interesting idea would be to additionally include a trending topic detection mechanism and automatically feed the resulting terms into the tracking visualization. Hereby, it would be possible to create a large landscape of events and topics.

References

[Bre12] BREWER C.: Color brewer 2.0. URL: http://www.colorbrewer2.org. 2

[SJ88] SPARCK JONES K.: *A statistical interpretation of term specificity and its application in retrieval*. Taylor Graham Publishing, 1988, pp. 132–142. 1

[TBP*10] THELWALL M., BUCKLEY K., PALTOGLOU G., CAI D., KAPPAS A.: Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (2010), 2544–2558. 1

[Twi] TWITTER TEAM: Developing for @twitterapi, 2013, https://dev.twitter.com. 2