Interactive Visual Analysis of High Throughput Text Streams

Chad A. Steed
Oak Ridge National Lab
Oak Ridge, TN, USA
csteed@acm.org

Thomas E. Potok
Oak Ridge National Lab
Oak Ridge, TN, USA
potokte@ornl.gov

Robert M. Patton
Oak Ridge National Lab
Oak Ridge, TN, USA
pattonrm@ornl.gov

John R. Goodall
Oak Ridge National Lab
Oak Ridge, TN, USA
jgoodall@ornl.gov

Christopher Maness
Oak Ridge National Lab
Oak Ridge, TN, USA
manesscs1@ornl.gov

James Senter
Cornell University
Ithaca, NY, USA
jksenter@gmail.com

ABSTRACT

The scale, velocity, and dynamic nature of large scale social media systems like Twitter demand a new set of visual analytics techniques that support near real-time situational awareness. Social media systems are credited with escalating social protest during recent large scale riots. Virtual communities form rapidly in these online systems, and they occasionally foster violence and unrest which is conveyed in the users' language. Techniques for analyzing broad trends over these networks or reconstructing conversations within small groups have been demonstrated in recent years, but state-of-the-art tools are inadequate at supporting near real-time analysis of these high throughput streams of unstructured information. In this paper, we present an adaptive system to discover and interactively explore these virtual networks, as well as detect sentiment, highlight change, and discover spatio-temporal patterns.

Author Keywords

Visual analytics; text analytics; stream analysis; visualization; sentiment analysis; change detection; social media; geospatial; temporal.

ACM Classification Keywords

I.2.7 Natural Language Processing: Text analysis

General Terms

Human Factors; Design.

INTRODUCTION

Notwithstanding their ubiquity, social media and other similar online streams of textual information represent a largely untapped resource of collective wisdom for in situ awareness and predictive understanding of global events. The rising value of these platforms is apparent as we see journalists increasingly turning to these rich sources of user content to track the importance of events and find sources of expertise. The streams are generally characterized by high throughput sources that produce large and rapidly changing content in the form of unstructured, textual information. If we consider events such as the London Riots of 2011–an infamous series of events where social media systems are credited with giving critical mass to the larger movement–it is evident that these systems are rapidly transforming public discourse and setting trends and agendas in areas such as politics, technology, and the environment [2, 1]. Effectively investigating such resources requires the ability to grasp the pulse of the network of information and drill-down to increasingly detailed perspectives in near real-time. The goal of such analysis is to reveal key connections, associations, and anomalies by harnessing a visual analytics approach that combines the strengths of humans (e.g. rapid visual pattern recognition, background knowledge, intuition) with the computational processing power of machines.

RELATED WORKS

The scale, velocity, and complexity of streaming content from social media and similar online feeds render state-of-the-art tools largely inadequate at supporting interactive analysis, thus, highlighting the novelty of the proposed research. A related system is STREAMIT, which is designed to explore streaming text documents via force directed graphs based on document similarities [1]. However, STREAMIT does not address the need for near real-time processing of high-throughput streaming content–the key objective of our research presented here. In the text-mining literature, we find either purely visual approaches that do not scale beyond moderate size collections [3, 4, 8, 9, 10], or purely automated approaches that restrict human interaction and carry significant trust issues with respect to the results [5, 6, 7]. Existing systems do not address the need for full spectrum, interactive analysis or the extension of said techniques to address dynamic sources and ever-changing analytical questions.

In light of said challenges, our research combines the flexibility, creativity, and domain expertise of humans with the tremendous computational and storage capacities of machines in a visual analytics framework for interactive analysis of high-throughput, unstructured information streams. This framework will facilitate in situ exploration of unstructured

The 2nd Workshop on Interactive Visual Text Analytics, Oct. 15, 2012, Seattle, WA, USA.

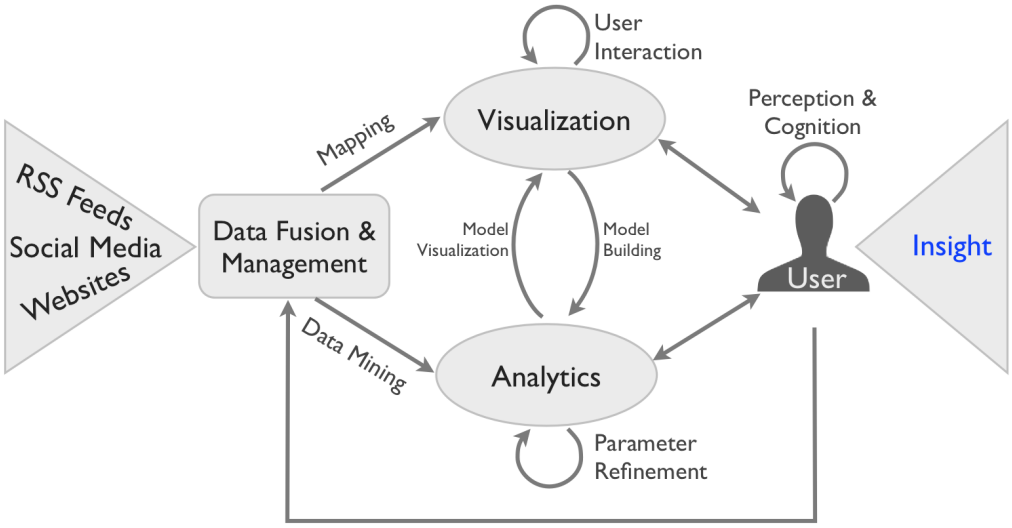


Figure 1. This notional diagram of our visual text analytics framework illustrates how streams of text are consumed and and fused for subsequent processing by automated analytics and interactive visualization techniques. Human interaction and feedback is incorporated into the system through a continuous feedback loop.

text streams, as well as detect sentiment, highlight change, and discover spatio-temporal patterns within the network. In the remainder of this paper, we will provide of overview of the proposed framework with descriptions of initial prototype components that are focused on geospatial and temporal exploratory analysis.

APPROACH

As shown in Figure 1, our visual analytics framework consists of data management, analytics, and visualization components that work together to transform unstructured text streams into insight. Since a discussion of the data management component is beyond the scope of the current work, we will focus on the visualization and analytics components. Initially, our interactive visualizations are focused on revealing specific insight regarding temporal and geospatial features in the streaming information.

Temporal Stream Visualizations

As shown in Figure 2, we have developed a temporal term frequency visualization technique for streaming information with support for interactive queries. The visualization technique also supports interactively browsing through the history of items for a time window of interest. In this view, the line graphs represent temporal term frequencies from a stream of news article summaries. The graph is animated so that as new items are processed, a new point is appended to the line plot, and the plot is translated to the left by one time unit. The vertical red line that spans the graphs shown in Figure 2 tracks the location of the mouse cursor to identify the frequency of term occurrences at the instant of time associated with the mouse location. The user can also use the mouse to drag a rectangular selection in the display to view all associated documents in the selected time range. Tick marks are shown beneath the graphs to highlight significant changes in term frequencies based on a user-defined threshold value. On the left-hand side of the graphs, the term of interest and maximum frequency value are shown. This stream term frequency visualization is designed to assist in the detection and investigation of short term atypical events as well as the discovery of longer term shifts in topical contents.

This temporal stream visualization is being extended to encode additional attributes of the data utilizing efficient data attribute to visual feature mappings such as connectedness, color, and shapes, to encode an appropriate number of dimensions into a single display. These visualizations will be augmented with information derived from stream data analytics so that significant associations and events are highlighted to increase visual saliency and improve the likelihood of discovering relevant insight for the subject of interest.

Change Detection and Sentiment Analysis

New algorithms to incrementally analyze a large number of streaming textual items in parallel are a key challenge for summarizing and highlighting potentially significant associations. Algorithms for efficient change detection and clustering of textual information will assist in the detection of significant events and the associations of items thereby improving the situational awareness in global overviews. Incremental sentiment analysis algorithms that facilitate an understanding of opinion, emotion, and subjectivity in text will be key metrics for characterizing the streams and guiding the analysis process to support spatio-temporal change detection tasks. From estimating stock prices based on anxiety in text to predicting the outcome of presidential debates, sentiment changes in social media streams has demonstrated much promise in predicting or associating with offline phenomena. We are developing algorithms to ascertain sentiment in a near real-time manner for important online events using an approach that combines machine learning, lexicon-based sentiment assignment, and linguistic analysis. Our approach includes various refinements, such as filtering for features and

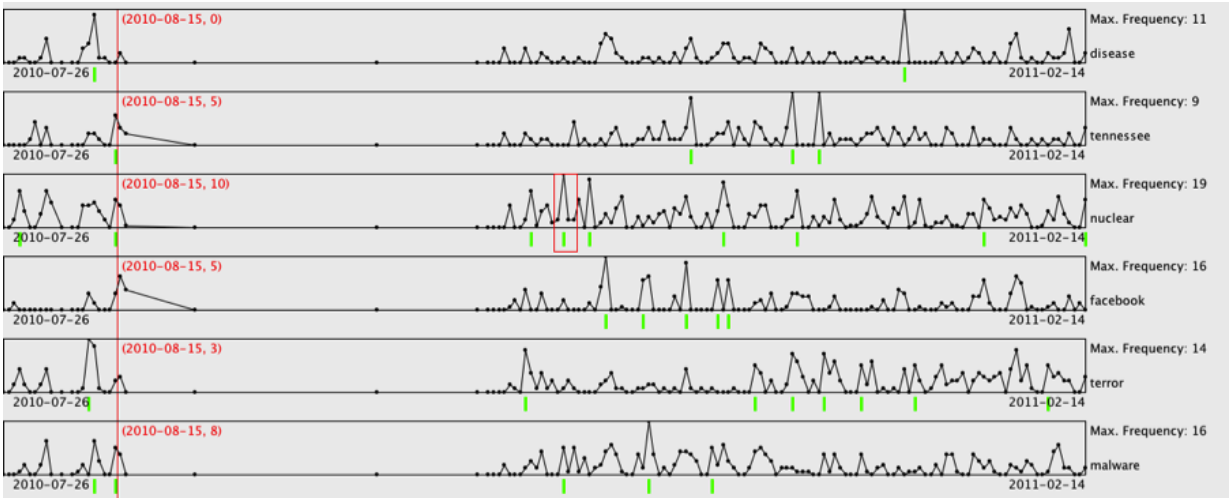


Figure 2. The stream visualization uses animated graphs to depict the temporal term frequency patterns in a text stream. In this case, 5 terms are being monitored from a stream of news article summaries. Tick marks are rendered beneath the graphs where the term frequency undergoes a significant change according to a user-defined threshold.

extensions to cope with the incremental addition of new items as well as utilization of geospatial features and entities to analyze geographical change in sentiment. The results of such analysis will then be used to augment the stream visualizations.

Geospatial Visualizations

The framework also leverages geospatial metadata from the stream to reveal temporal trends for specific areas of the world. Information from these algorithms facilitate a spatio-temporal awareness which can be ascertained via geovisualizations such as our heatmap visualization of Twitter updates from the opening weekend of the 2012 London Olympics (see Figure 3). In this image, the map is treated as a set of equally spaced bins which accumulate tweets based on associated geographical metadata. The count of tweets for each bin is then encoded with a blue color scale such that areas shaded with highly saturated blue are indicative of the highest tweet counts. For example, in Figure 3(a), the highest concentration of tweets occur at the Olympic Stadium because the time range covers the opening ceremony event held there. By contrast, the heatmap in Figure 3(b) shows tweets 48 hours after the ceremony. The tweets are more dispersed since multiple events occurred throughout the region.

Intelligent User Interfaces

A human user is expected to see key associations and trends in the data and also create and refine queries and parameters for both the analytics and visualization components of the system. To support such functionality, the system will assist users in the analysis process by adapting the user interface using semi-supervised machine learning and pattern recognition techniques. As the application tracks interactions with the visualizations and graphical display widgets, the user will be able to visually create and refine analytical questions that drive the parameters of the analytics algorithms. For example, given a clustering of items for a topic of interest, the users interactions with the results are recorded and used to label documents as relevant or irrelevant. These labeled items can then be examined programmatically to re-display the remaining unlabeled items in a manner that increases the prominence of potentially relevant items, thereby increasing the likelihood of finding such information that may be buried in the more obscure portions of the display.

CONCLUSION

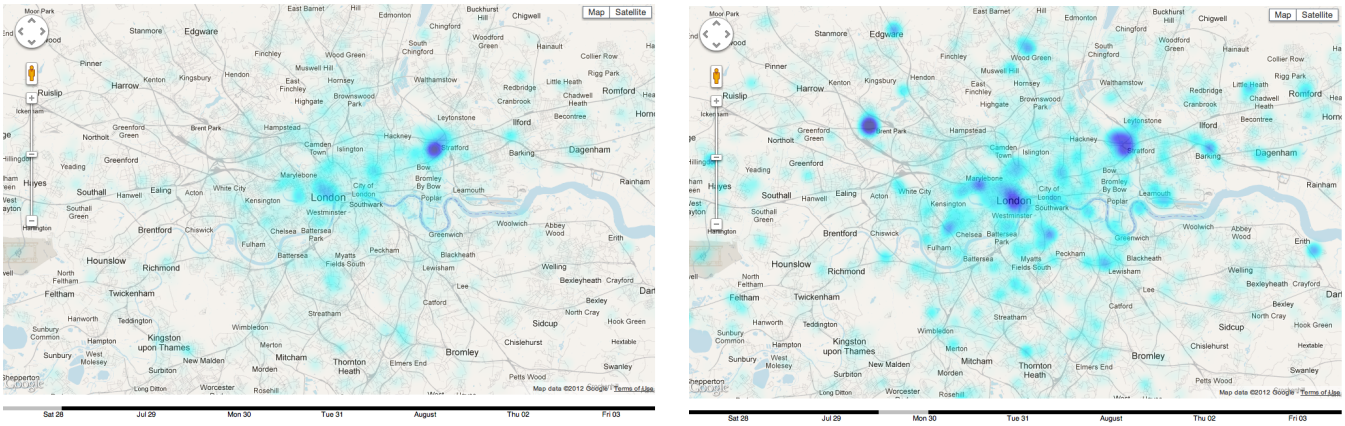
In the current work, we provide an overview of our research on interactive exploration of high-throughput, unstructured text streams. Additionally, details on term frequency visualizations and temporal geospatial visualizations have been discussed. We are currently extending this visual analytics framework to integrate additional data from the raw information stream, new interactive exploratory capabilities, information from stream-based analytics, and semi-supervised machine learning components that adapt the display based on user interactions.

ACKNOWLEDGMENTS

This research is sponsored by the Oak Ridge National Laboratory's Laboratory Directed Research and Development (LDRD) Fund. This paper was prepared by the Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT–Battelle, LLC, for the U.S. Department of Energy, under contract DE-AC05-00OR22725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

1. Asur, S., and Huberman, B. A. Predicting the future with social media. *CoRR abs/1003.5699* (2010).
2. Bohannon, J. Tweeting the london riots. *Science 336*, 6083 (2012), 831.



(a) Night of Opening Ceremony

(b) Two Days After Opening Ceremony

Figure 3. We generate heatmaps using a blue color scale based on the geographical metadata associated with Twitter status updates. The more saturated blue areas are indicative of a higher concentration of tweets. In this case, we show Twitter activity during the first weekend of the 2012 London Olympics for (a) the night of the opening ceremony and (b) two days after the opening ceremony. Tweets are clustered around the Olympic Stadium in (a), but are more dispersed in (b) since events occur at multiple venues throughout the region. We use the Google Maps API to generate these heatmaps.

3. Cao, N., Sun, J., Lin, Y., Gotz, D., Liu, S., and Qu, H. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Trans. on Visualization and Computer Graphics 16*, 6 (2010), 1172–1181.
4. Dörk, M., Carpendale, S., Collins, C., and Williamson, C. Visgets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Trans. on Visualization and Computer Graphics 14*, 6 (2008), 1205–1212.
5. Frunza, O., Inkpen, D., and Tran, T. A machine learning approach for identifying disease-treatment relations in short texts. *IEEE Trans. on Knowledge and Data Engineering* (2011).
6. Keim, D. A., Mansmann, F., and Thomas, J. Visual analytics: how much visualization and how much analytics? *SIGKDD Explorations Newsl. 11* (May 2010), 5–8.
7. Lin, C., He, Y., Everson, R., and Röger, S. Weakly-supervised joint sentiment-topic detection from text. *IEEE Trans. on Knowledge and Data Engineering* (2011).
8. Stasko, J., Görg, C., and Liu, Z. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization 7*, 2 (2008), 118–132.
9. Strobelt, H., Oelke, D., Rohrdantz, C., Stoffel, A., Keim, D. A., and Deussen, O. Document cards: A top trumps visualization for documents. *IEEE Trans. on Visualization and Computer Graphics 15*, 6 (2009), 1145–1152.
10. Wong, P. C., Hetzler, B., Posse, C., Whiting, M., Harve, S., Cramer, N., Shah, A., Singhal, M., Turner, A., and Thomas, J. In-spire infovis 2004 contest entry. In *IEEE Symposium on Information Visualization* (Oct. 2004).