TextPool: Visualizing Live Text Streams

Conrad Albrecht-Buehler*
Northwestern University

Benjamin Watson*
Northwestern University

David A. Shamma*
Northwestern University

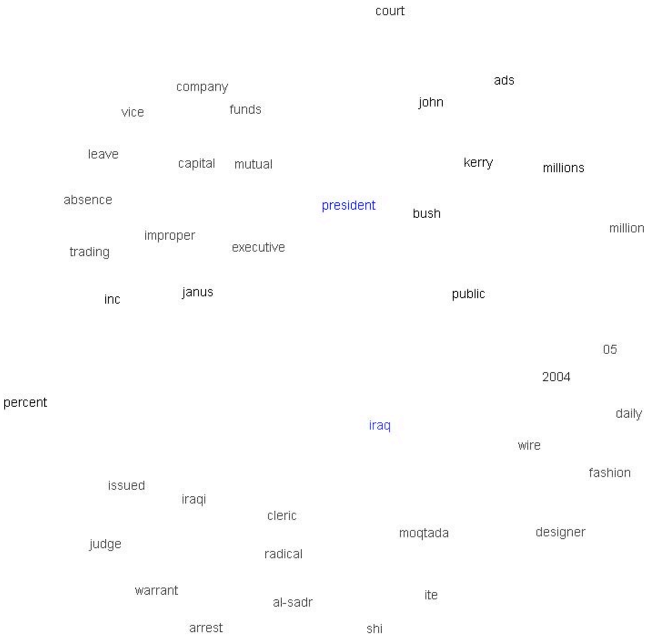

Figure 1: A *TextPool* visualization of six hours of content in several news feeds on Monday, April 5, 2004. Here the user has focused on stories related to the terms “president” and “iraq”. This focused display is derived from 137 stories as represented by 1,662 terms. Note the discussion of improper trading, an Iraqi cleric, and fashion.

Abstract

In today's fast-paced world it is becoming increasingly difficult to stay abreast of the public discourse. With the advent of hundreds of closed-captioned cable channels and internet-based channels such as news feeds, blogs, or email, knowing the "buzz" is a particular challenge. *TextPool* addresses this problem by quickly summarizing recent content in live text streams. The summarization is a dynamically changing textual collage that clusters related terms. We tested *TextPool* with the content of several RSS newswire feeds, which are updated roughly every five minutes. *TextPool* was able to handle this bandwidth well, and produced useful summarizations of feed content.

CR Categories: I.3.8 [Computing Methodologies]: Computer Graphics – Applications; I.7.0 [Computing Methodologies]: Document and Text Processing – General

Keywords: information visualization, data streams, text layout, information retrieval, newswires.

1 Visualizing Live Text Streams

TextPool is a tool for visualizing and maintaining an up-to-the-minute understanding of live streams of text such as newswires and closed-captioned television. Previous work has focused on providing a topic space for daily news content [3], overview visualizations of static [5,8] or dynamic [9] corpora, or created maps of thematic change in a corpus [1,2]. *TextPool* is a "buzz" visualizer, providing continuous access to the prominent subjects of discourse within one or more streams. *TextPool* buffers and processes text streams in real time using information retrieval (IR) techniques, extracts the most significant terms from the buffered streams, and displays related terms in proximity to one another in a text collage (Figure 1) that is adjusted dynamically in response to user interaction and changes in stream content.

2 Reading Live Text Streams

Our goal was to build a tool for visualizing text streams, including news feeds, closed captioning, email, and blogs. We chose news stories as our test dataset, because they are published frequently and represent an aggregate from a broad range of sources, such as wire services, newspapers, or local affiliates. We monitor and log news stories published as Really Simple Syndication (RSS) feeds [4] by Yahoo! News in a database that acts as a buffer of recent content independent of the text visualizer. The visualization client then retrieves the latest feed data from the buffer.

We represent the news stories by creating content vectors from their titles and the 10-30 word descriptions associated with them as part of the RSS feed. Previously Shamma et. al. observed that these descriptions are adequate representations of the content of each RSS story [7], eliminating the need to find content vectors using inverse document frequency [6].

As a measure of salience across multiple stories and streams, we create a co-occurrence matrix. We rely on the fact that words that are used together likely have meaning together, and that if terms co-occur in several stories, then they are also more representative of the current discourse.

3 Dynamic Visualization

TextPool is designed to convey the relatedness of terms by their proximity to one another in the display. To accomplish this, we present a graph in which nodes represent salient terms from the stream, and are connected to their co-occurring terms by connections whose lengths are scaled by the inverse of their co-occurrence, so that terms that are closely related and co-occur often are close to one another in the graph. By simulating the term nodes as masses, and their interconnections as damped springs, we can lay out the display in real-time with the added advantage of enabling the user to see how the display rearranges over time (Figure 2).

4 Supported Interactions

Controlling temporal context. Users can control the temporal context of the information displayed using a temporal window that indicates how much of the recent stream should be visualized. As the stream moves through the window, old news items that have moved beyond the window are removed from the graph, and

*{conrad,watsonb,ayman}@cs.northwestern.edu

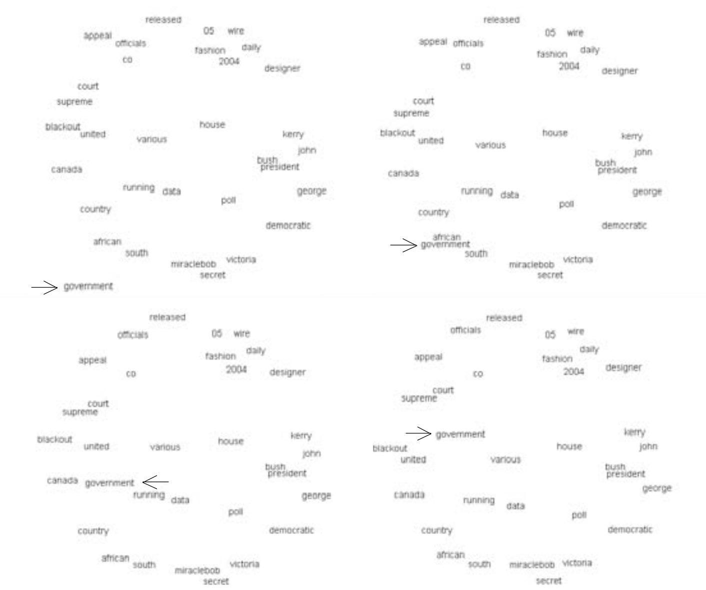

Figure 2: Some new stories arrive in the stream, making the term "government" salient enough for display. The term arrives at the lower left, and gradually works its way into the current visualization, finding an optimal position next to the terms "supreme" and "court". (Arrow added for illustrative purposes and visualization frames are ordered left to right, top to bottom)

those that have just been published and entered the window are added to the graph.

Focusing and zooming. In addition, by selecting several displayed terms, users can limit display to those terms and any that co-occur with them. *TextPool* then allows users to zoom in on the data-space with a slider that can reduce the minimum frequency of co-occurrence of displayed terms, revealing lower volume discourse related to the selected terms. *TextPool* also permits display space zooming with a slider that scales the lengths of all links. At any point, *TextPool* can also provide the user with direct access to the documents containing the currently displayed terms, allowing closer analysis.

Highlighting. To help identify hot topics that are the focus of recent discussion, *TextPool* brightens more recently discussed terms and dims older terms. By default, node interconnections are not displayed, but *TextPool* users can reveal them, highlighting term relatedness (Figure 3). This is particularly useful when the display is densely populated. Users can also highlight relatedness through motion by temporarily adding energy to the spring-mass system. In the resulting ripple, related groups of terms move in a coordinated fashion.

5 Future Work

In the near term, we will be increasing *TextPool*'s capacity so that it can handle even higher bandwidth streams and create visualizations with larger temporal contexts. Although *TextPool* supports two very different patterns of use, one passive and the other active, it has not been specialized for either of those uses. We plan to examine the possibility of such specialization. Looking beyond news stories, there are many other forms of streaming text available to us - we plan to examine email, blogs, and closed captioning as well. We will also experiment with additional display dimensions such as text size, color and motion, subject to the understandability concerns raised by graphic design.

6 Conclusion

TextPool is an interactive system that summarizes the latest content in live text streams. *TextPool*'s visual summary is a dynamically changing textual collage, in which co-occurring terms are grouped together. *TextPool* can be used in an ambient mode, offering a peripheral awareness of stream traffic; or an interactive mode, supporting a deeper understanding of stream content with focus and zoom functionality. We tested *TextPool* by using it to visualize the content of several RSS news feeds at one time. *TextPool* was able to handle this volume well, and produced useful summarizations of current RSS feed content.

7 Acknowledgements

Our thanks to Kris Hammond, Larry Birnbaum and the InfoLab for their suggestions and advice, and to Proctor and Gamble for their encouragement and interest. This research was funded by NSF Award 0093172.

8 References

1. S. Havre, B. Hetzler & L. Nowell. 2000. Proc. IEEE Information Visualization, 115-123.
2. N. Miller, P. Wong, M. Brewster & H. Foote. 1998. TOPIC ISLANDS – a wavelet-based text visualization system. Proc. IEEE Visualization, 189-196.
3. E. Rennison. 1994. Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes. Seventh Annual Symposium on User Interface Software and Technology, 3-12.
4. RSS Advisory Board. 2003. RSS 2.0 specification. http://blogs.law.harvard.edu/tech/rss.
5. D. Rushall & M. Ilgen. 1996. DEPICT: documents evaluated as pictures. Proc. IEEE Information Visualization, 100-107.
6. G. Salton, A. Wong & C. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, *18*, 613-620.
7. D. A. Shamma, S. Owsley, K. J. Hammond, S. Bradshaw, and J. Budzik. Network Arts: Exposing cultural reality. In *Alternate track papers & posters of the 13th international conference on World Wide Web*, pages 41–47. ACM Press, 2004.
8. J. Wise, J. Thmoas, K. Pennock, D. Lantrip, M. Pottier, A. Schur & V. Crow. 1995. Visualizing the non-visual: spatial analysis and interaction with information from text documents. Proc. IEEE Information Visualization, 51-58.
9. P. Wong, H. Foote, D. Adams, W. Cowley & J. Thomas. 2003. Dynamic visualization of transient data streams. Proc. IEEE Information Visualization, 97-104.

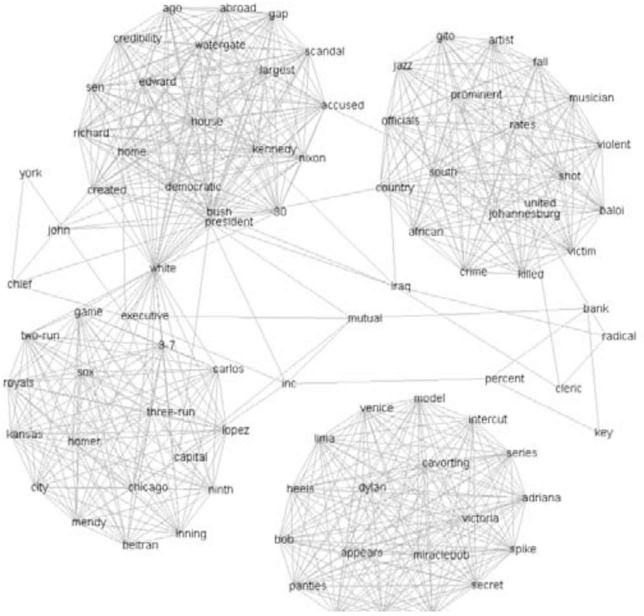

Figure 3: In *TextPool*, nodes represent salient terms from the stream, and are connected if they co-occur within a story. Connection length shortens as frequency of co-occurrence increases, revealing term relatedness.