BreakingStory: Visualizing Change in Online News

Jean Anne Fitzpatrick
James Reffell
Moryma Aydelott
School of Information Management and Systems
University of California at Berkeley
Berkeley CA 94720 USA
+1 510 642 1464
{jafitz, reffell, aydelott}@alumni.sims.berkeley.edu

ABSTRACT

BreakingStory is an interactive system for visualizing change in online news. The system regularly collects the text from the front pages of international daily news web sites. It allows users to search over the collection and view the frequency of occurrence of keywords in graphic, tabular, and full text formats. Results from the system are shown over time, and can be filtered geographically. The system was developed using a user-centered design process that included rapid prototyping and informal user testing. It provides a new way of viewing the news that incorporates a sense of history.

Keywords

visualization, news analysis, WWW, search user interfaces, iterative design

OVERVIEW

News as a genre has always been crisis-driven, focusing on today's "breaking story". With few exceptions, even events of previous days get only cursory reference, and broader historical context is completely lacking. As news moves to the Internet, the potential exists for an ever more rapid and continuous cycle of change, and even less preservation of historical context: as a story breaks, news web pages are updated and the previous versions cease to exist.

The Internet and digital archives also have the potential to reverse this trend, providing ready access to news stories over time and from a range of sources. However, while there are many systems for searching news text (including commercial services such as Lexis-Nexis and on-line search engines), they generally provide only a relevancy-ranked list of search results. While useful for finding specific documents of interest, the temporal context of the document is not highlighted.

The Internet also has the potential to provide access to news sources from around the world, exposing news consumers to multiple points of view on a given story. News sources in different geographic regions can demonstrate different points of view in terms of both story selection and presentation[1]. However, existing systems select sources based on popularity or "quality", rather than helping to find related stories across geographically diverse news sources.

BreakingStory was developed as one possible solution to these problems. The system allows users to search for words and phrases that appear on the front pages of a wide range of online news sites. The results of the search are shown graphically as trends over time, in tables of summary data, and in the actual page text. Users can explore the data across date ranges and geographic regions and can compare news sites and regions by viewing multiple charts simultaneously.

Our system draws upon prior research in news analysis, including the general issues of point of view (especially as related to geography)[1,2,3] and context (especially historical context)[4,5], with a particular interest in prior attempts to visualize news data[6,7]. It provides a direct visual representation of term frequency, taking advantage of human perceptual and cognitive abilities for pattern recognition[8]. The system represents a unique combination of capabilities that have not previously been applied to a news corpus.

SYSTEM AND DATA

The BreakingStory system provides search capabilities for a corpus of news web site front pages, with display capabilities including aggregation and filtering by date range and by geographic affiliation of the news sites. The system is implemented in Java and Perl, with a MySQL database for hierarchical geographic metadata. Indexing and retrieval functions are based on Lucene, an open source search engine.[1] Modifications were made to adapt Lucene for BreakingStory; most importantly, the TF/IDF scoring was overridden to return raw term frequencies.

BreakingStory's core dataset is comprised of the full text of the front pages of more than 100 news web sites in over 50 countries around the world. While limited to English language sources that update daily, we have attempted to provide broad geographic coverage, with at least one site in

CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-630-7/03/0004.

[1] http://jakarta.apache.org/lucene/docs/index.html

every geographic subregion.[2] Each page is associated with temporal and geographic metadata, allowing us to filter and aggregate search results by both facets simultaneously.

We deliberately chose to collect only the news web site front pages, rather than spidering the sites to obtain all of their content. The front page contains the news items considered most important by site editors, and is therefore most indicative of the editorial point of view. Collecting the front pages helps to compensate for the broad use of wire services: while sites may carry many of the same stories, they may not always place them on the front page. A third advantage is that it provides rough normalization for the widely varying size of web sites. Finally, front pages are updated regularly, highlighting changes over time.

VISUALIZATION

BreakingStory's users visualize changes in news by navigating through a flow of simple views as they explore the dataset. Large line charts show the details of a news source or collection of news sources, while small multiples of the same charts provide for outlier detection and comparisons across groups.

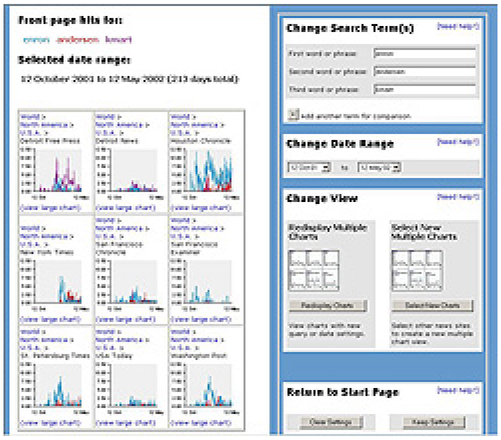

Figure 1: Small multiples for visual comparison.

Below the charts, we provide additional context in the form of detailed query preview tables that display geographic and temporal breakdowns of the overall values. We also provide access to the full text of the returned front pages with the search terms and surrounding sentences highlighted for easy visual parsing.

DESIGN METHODOLOGY

Our development approach was centered on an iterative, user-centered design process. The interaction design for BreakingStory began with interviews of members of our target user groups: professional media critics and analysts and highly motivated non-professional media analysts. We rapidly iterated through paper and interactive prototypes, evaluating each with informal user tests. The testing focused on information design as well as interaction design, because task success depended on accurate interpretation of the charts. Each iteration in testing resulted in improvements to both the visualization and the interaction model.

We have identified many areas for future work in the user interface, the underlying system, and the data set. Potential avenues include more sophisticated visualizations, expanded search capabilities, natural language processing, and increased breadth or depth of corpus.

CONCLUSION

Using simple but effective graphical and textual views of a novel data set combined with a streamlined and flexible interaction design, BreakingStory provides a possible new paradigm for exploring and visualizing temporally organized textual data. It can be used as an analysis tool itself and as a starting point from which other kinds of media analysis can be initiated. BreakingStory allows analysts and more general news audiences to approach complex news data in an exploratory fashion, interactively forming and testing hypotheses to achieve both a big picture understanding of changes in news coverage and a detailed view of the coverage that most interests them. It offers wider temporal context for relating to the news, and will hopefully contribute to imbuing the news with a sense of history.

REFERENCES

1. Dijk, T.A. van (1988). News Analysis: Case Studies of International and National News in the Press. Hillsdale: Erlbaum.
2. Sack, W. (1995). Representing and Recognizing Point of View. The American Association of Artificial Intelligence Fall Symposium on Artificial Intelligence Applications in Knowledge Navigation and Retrieval. Cambridge, MA: AAAI Press.
3. Elo, S. PLUM: Contextualizing News for Communities Through Augmentation.(1995). Cambridge, MA.
4. Hodge, G. and Kress, R. Language as Ideology (Politics of Language). (1979). London / New York: Routledge.
5. Sack, W. The Questioning News System (Unpublished Working Paper). (1997). Cambridge, MA: News of the Future Consortium, MIT Media Laboratory.
6. Wise, J. E., Thomas, J. J., Pennock, K. Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. Proceedings of InfoVis'95, IEEE Symposium on Information Visualization, New York.
7. Rennison, E. (1994). Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes. Proceedings of UIST'94, ACM Symposium on User Interface Software and Technology, New York. 3-12.
8. Kosslyn, S. M., Pinker, S. Simcox, W. and Parkin, L. (1983). Understanding Charts and Graphs: A Project in Applied Cognitive Science. NIE 83 ED 1.310/2:238687.

[2] Geographic data adapted from the United Nations Statistics Division (http://www.un.org/Depts/unsd/).