

Web-Based Visual Analytics for Social Media

Daniel M. Best, Joe Bruce, Scott Dowson, Oriana Love, and Liam McGrath

Pacific Northwest National Laboratory
902 Battelle Boulevard
Richland, WA 99352 USA

Abstract

Social media provides a rich source of data that reflects current trends on a multitude of topics. The data can be harvested from Twitter, Facebook, blogs, and other social applications. The high rate of adoption of social media has created a domain that is difficult to analyze, due to the ever-expanding volume of data. Information visualization is key in drawing out features of interest in social media. The Scalable Reasoning System is an application that couples a back-end server equipped with analysis algorithms and an intuitive visual interface to allow for investigation. We provide a componentized system that can be rapidly adapted to user needs. The information in which they are most interested is featured prominently in the application. As an example, we have developed a weather and traffic monitoring application for use by emergency operators in the city of Seattle.

1 Introduction

Social media continues to grow as an outlet for the public to express their feelings toward many subjects of daily life. The ability to draw out features of interest is important for citizens, private business and government entities. Private business can use information provided by social media to fine tune advertising campaigns and other financial planning decisions. For example, if a particular feature of a product is trending high with young adults then the company can update advertisements to that demographic. Government entities can utilize the data made available by gaining a general awareness of their area of oversight. Icy sections of road can potentially be identified through social media (e.g., Twitter) in addition to traditional means.

The amount of data produced by social media feeds from Twitter and Facebook (hundreds of tweets and statuses per second) is staggering and difficult for an individual to process. To adequately harness the power of social media, analysis algorithms and visualization are needed to aid in trending and pattern discovery. To help the investigation process, an intuitive display assists by linking visualization and data together. The underlying data linked with the visualizations provides a detail-on-demand interaction, which is important for knowledge generation.

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The Scalable Reasoning System (SRS) is just such an application. We have developed SRS in a way that allows for customization of the interface based on end-user needs and visualizations that best fit the data being investigated. The visual metaphors and back-end algorithms utilized by SRS provide the ability for impromptu analytics for a wide range of data types including social media. Social media presents a unique set of challenges: the brevity of most messages, the frequency of data ingest, and the context sensitivity of each message. SRS has addressed several of these issues and adapted visualizations to best suit social media investigation.

2 Related Work

Crisis informatics is quickly being defined as a new sub-field within informatics as more data – such as Twitter – is available to emergency providers. Like many other branches of government, the Department of Homeland Security is working toward establishing a social media strategy as they recognize that citizens and emergency personnel can communicate within these mediums (DHS 2012). In Japan, where earthquakes are numerous, researchers are able to provide earthquake alerts to registered users faster than the Japan Meteorological Agency (Sakaki, Okazaki, and Matsuo 2010). By looking at four major events – two of which involved Category 4 hurricanes – researchers considered how users communicated differently within Twitter (Hughes and Palen 2009). They found that, compared to common tweets, event tweets contained more information, were passed more effectively between people, were replied to less often, and contained more URLs.

Previous publications show that governments have embraced Twitter in its own ways. Taking a theoretical approach toward the adoption of Twitter, Wigand et al (Wigand 2010) explore how unexpected groups like government have re-appropriated the microblogging tool to reach out to the public sector, communicate with other branches of government, build relationships and collaborate. Likewise, after manually coding several hundred Tweets from 69 members of congress, Golbeck et al (Golbeck, Grimes, and Rogers 2010) found that most tweets were informational in nature, few were official, and only 7.4% were intended for people outside of congress.

Aggregate visualizations allow users to reduce a corpus of tweets into high level summaries. The TweetInfo visu-

alization identifies subevents along a timeline while also summarizing the sentiment expressed in the topical tweets (Marcus et al. 2011). By making the timeline the central element of the visualization, TweetInfo encourages the identification of voluminous data events, though users can easily drill down for a more fine-grained view. This real-time Twitter analysis capability is a keen way to help with major event detection while also suggesting sentiment. The Seriously Rapid Source Review highlights interesting and trustworthy sources for journalists within a sea of soft and hard news (Diakopoulos, De Choudhury, and Naaman 2012). The user interface supports user role identification (i.e., journalist/blogger, organization, ordinary person), location, top entities mentioned and marks probable witnesses to aid journalist in finding and assessing sources.

3 Scalable Reasoning System

SRS is a modular framework that allows rapid prototyping of visual analytic applications through sound componentization. A back-end component is responsible for interfacing with data stores, algorithms, and general processing needs to support the analytic process. The web front-end displays the results of the algorithms in a way that users can understand and interact with to investigate their information space. Both front-end and back-end can be customized to accommodate the needs of a particular deployment. The Customization of SRS does take development effort, as little as an hour for minimal changes (layout, use of a different widget, etc) to full development activities to support new visualizations and data types.

City of Seattle and Twitter

A user-centric approach was taken to understand the needs of the City of Seattle’s Emergency Operations Center (EOC) personnel and build a visualization that would prove useful. The city’s EOC is activated eight times a year on average typically due to weather related events. During these incidents, Police, Fire, City Light (Power), Seattle Public Utilities (Water), Department of Transportation, Health Department, the Mayors office, and other emergency personnel work together to provide critical services and communicate with impacted citizens. Depending on the situation, the EOC may aim to restore power, address road closures or severe accidents, or even supply food, water, and shelter.

The City of Seattle actively uses Twitter today, mainly to communicate with the public by broadcasting messages. We engaged the city through focus groups and one-on-one sessions to draw out the primary use cases and design our application. The city was interested in understanding monitoring the reach of critical messages and in general, gaining a better and more timely understanding of situations where the EOC is activated. Based on survey results, most already had a positive impression of social media citing Twitter as being *timely*, *useful* and *convenient*. Our challenge is to support their use cases while maintaining these positive impressions of social media within a visualization tool.

Back-end

The back-end framework is the set of deployed components that determine the available data and analytic algorithms for a custom application. To support a variety of deployments, it is possible to provide security extensions to secure data, algorithmic plug-ins to analyze the data, data extension components to supplement the feeds, and data adapters to communicate with external data-stores.

For the City of Seattle, a two-tier architecture was designed where Twitter data is constantly harvested based on a user’s set of feeds using SmartFeeds, an internally developed content harvester based on Drupal (Byron, Robbins, and Berry 2009), and stored in a database. This approach provides both storage of historic tweets while also providing access to newly published tweets within minutes. Furthermore, SmartFeeds identifies redundant feeds, created when multiple users are using the same search terms, to reduce network bandwidth usage, storage requirements, and calls to the Twitter API.

The set of analytic capabilities available for an SRS deployment are determined by the analytic plugins utilized. These components encapsulate specific algorithms first to model a set of data, and then to provide the resulting output to the front-end web application, driving many of the visualizations. For the City of Seattle, this included plugins to analyze and create logical bins to organize tweets for a histogram visualization. To support visualizing themes, a plugin wrapping the RAKE (Rose et al. 2010) algorithm was used that identifies and extracts major topical themes from the tweets. Finally the ability to “find related” tweets when a user finds a particularly interesting tweet is supported by an analytic plugin that contains a modified term frequency-inverse document frequency algorithm tailored for short documents.

Front-end

The client framework is separated into three layers, much like the model-view-controller design pattern: the server interaction layer (model), the visualization layer (view), and the transform layer (controller). Each application built on this framework utilizes some of the visualizations available from the widget library and connects to one or more data sources made available by the server. For each application, new code is written in a transform layer which glues the independent server and visualization components together by transforming the server data into a structure acceptable to the visualization. The City of Seattle application utilizes the Storyflow, Histogram, List, and Facets widgets as seen in Figure 1. The application is delivered as a web application, built on web-standard HTML, CSS, and Javascript, with SVG for more complicated rendering.

Filter In most SRS applications, the widgets share the same data source, rendering the data in different ways to give the user many angles to explore a dataset. Filter operations limit the volume and variability of the data, narrowing to targets of interest. They can be performed by specifying conditions on a set of properties in a web form. Each property has a data type (e.g., string, number, geolocation) and

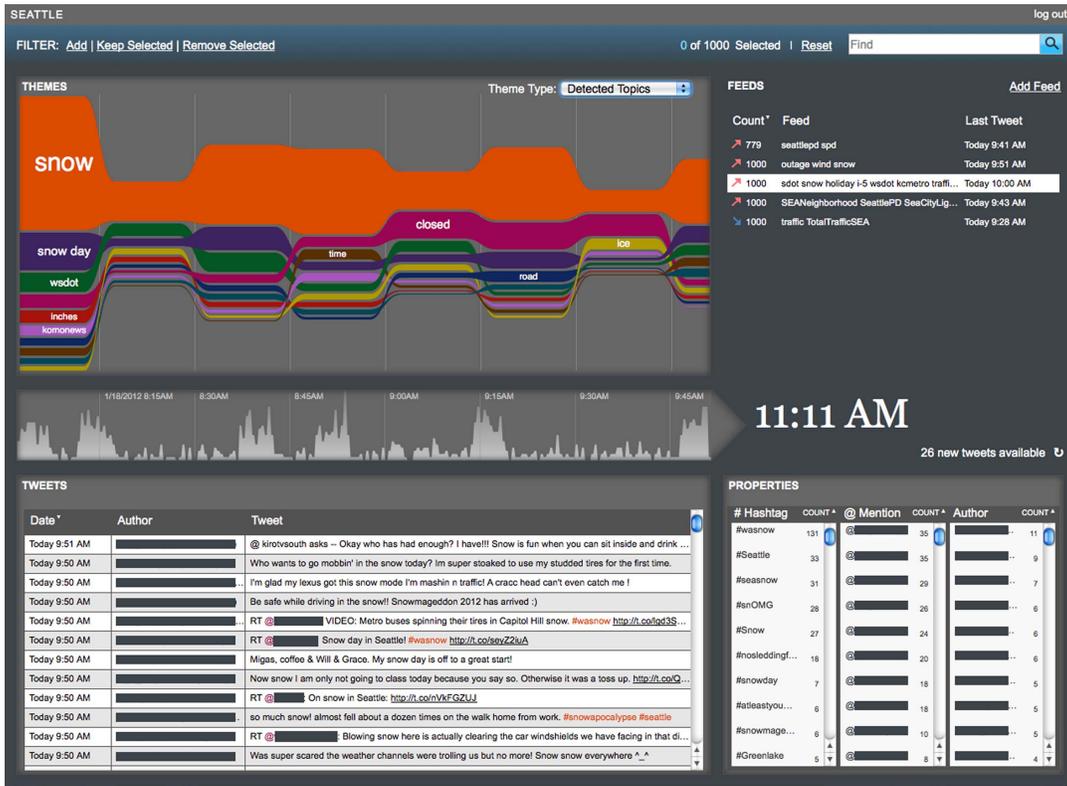


Figure 1: Scalable Reasoning System Social Media front-end.

appropriate operators (e.g. “contains”, less-than, “near”). A filter is one property paired with one operator and a value for that operator to use (i.e. a constant). They are restricted to singular properties for ease-of-comprehension, but they can also be chained to construct more complicated filters.

Brushing & Linking Filter operations can also be performed by building a selection set and subsetting to that set or its complement. Selection operations are performed in the widgets. Each glyph (visual element) in a widget is selectable, representing a subset of the dataset. The widgets render the data differently, so each selection interaction is different. In the histogram, a click-and-drag will select a time region. In the facets widget, clicking a facet will select all documents belonging to the facet. The widgets share a selection model, and it is linked across them, so they respond to each other’s selection operations. Glyphs can be rendered partially selected when appropriate. This allows the user to ascertain correlations across multiple dimensions in the data as they recognize patterns in the selection set.

Themes A streamgraph visualization (Byron and Wattenberg 2008) was chosen as the primary widget. It is capable of showing trends in the data over time. We employed Storyflow (Rose et al. 2009), which emphasizes the interchange of prominent themes, over a ThemeRiver (Havre, Hetzler, and Nowell 2000) -style streamgraph. The temporal binning and vertical story ordering would focus attention on the macro themes of the data as they fluctuated in prominence.

The top themes over the entire timeframe- rather than within each temporal bin- are displayed to provide continuity of themes across time. A negative side effect is leaving some time periods under represented, depicting those regions as largely void, though the themes there are merely less prominent. Each theme was assigned a color with the most prevalent themes granted a more pleasant color to maintain a welcoming aesthetic.

Five thematic viewing options support varying tasks. The “Top Hashtags” and “Top Mentions” options take advantage of the Twitter message structure. For the City of Seattle, this allows users to explore the fluctuations emergency themes. The “Top Topics” shows extracted prominent topics in the dataset. Feed search terms are taken from the terms the user provided to construct the feed; these may be hashtags, mentions, locations, or keywords. Finally, provided terms allow for further *ad-hoc* querying of the active dataset, building the storyflow visualization from arbitrary hashtags, mentions, and keywords provided by the user.

Real-Time Visualization The real-time nature of social media analytics for emergency management poses interesting visualization challenges. After meeting with the City of Seattle, we learned that “timeliness” was of the essence when considering Twitter data. As such, emphasizing this trait while maintaining a visualization that consistently evolved rather than constantly changed was necessary to avoid disorientation of the user.

Many visualizations present a static corpus of data that has been imported, processed and then translated into a visual representation. The dynamic and ever-changing nature of social media does not afford such algorithmic luxuries. Because our visualization diverged from the traditional pre-processed visualization model, we consciously decided to emphasize time by including a live clock in the visualization. The introduction of a clock may initially appear to be a minor and trite contribution, yet this obvious reminder helps users realize that visualization is synchronized with time just as the Twitter data is as well. As new tweets are being received, the refresh indicator next to the clock enables the user to draw in new data, preventing the visualization from changing in the midst of their exploration. The refresh indicator is accompanied by the number of new tweets available. This helps the user to determine if and when to incorporate new data. As only a portion of the timeframe is viewable based on the Twitter feed – the most recent 3 hours for example – some of the oldest tweets may be removed to incorporate the new tweets.

After feeds are added, they will immediately have a trend indicator which reveals the increasing or decreasing popularity of the terms within the feed. This indicator allows users to quickly access the “hottest” topics among their feed set.

4 Future Work

Future work for SRS is driven by feedback obtained from the City of Seattle after pilot deployments. One desire expressed by users is to utilize social media from sources in addition to Twitter, such as neighborhood blogs, so mechanisms for generalized harvesting and parsing of sources must be integrated.

Event detection (for alerting of real-life events of interest) and sentiment analysis (for judging public reaction to events) are also needed. With a large and growing number of sources, there is a need for additional filters.

A visualization that is utilized in other SRS applications is a geospatial representation of the document set. Due to the limitations of geolocation lookup for Twitter data, the map widget was not incorporated into the initial social media application. We envision utilizing geospatial visualization with a confidence associated with each marker. Markers location confidence may be determined from provided meta data (not often given), profile based information, gazetteer lookups, or content based analysis (Cheng, Caverlee, and Lee 2010).

5 Conclusion

Social media applications are on the rise due to the wealth of data that can be drawn from feeds. The content of social media can be used to provide insight into trends for a varying range of applications. The outcome of our implementation has been to assist Seattle EOC staff to better understand trends relating to topics that pertain to EOC operation. The combination of visual analytics in a lightweight web environment coupled with server side algorithmic support has provided a system that can be used by varying levels of staff to provide situational awareness and assist in the deci-

sion making process. Our test group of users found the tool *approachable* and *likely valuable* to their needs during an event. We will continue working with them to improve the interface as mentioned in Section 4 so that during an emergency they will have analytic capability to provide insight into public needs and concerns.

References

- Byron, L., and Wattenberg, M. 2008. Stacked graphs – geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics* 14:1245–1252.
- Byron, A.; Robbins, J.; and Berry, A. 2009. *Using Drupal*. O’Reilly Series. O’Reilly.
- Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 759–768. ACM.
2012. First responder communities of practice virtual social media working group social media strategy. Technical report, Department of Homeland Security Science and Technology Directorate.
- Diakopoulos, N.; De Choudhury, M.; and Naaman, M. 2012. Finding and assessing social media information sources in the context of journalism. preprint.
- Golbeck, J.; Grimes, J. M.; and Rogers, A. 2010. Twitter use by the u.s. congress. *J. Am. Soc. Inf. Sci. Technol.* 61:1612–1621.
- Havre, S.; Hetzler, B.; and Nowell, L. 2000. Themeriver: Visualizing theme changes over time. In *Proc. IEEE Symposium on Information Visualization*, 115–123.
- Hughes, A. L., and Palen, L. 2009. Twitter Adoption and Use in Mass Convergence and Emergency Events. In *IS-CRAM Conference*.
- Marcus, A.; Bernstein, M. S.; Badar, O.; Karger, D. R.; Maden, S.; and Miller, R. C. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI ’11, 227–236. New York, NY, USA: ACM.
- Rose, S.; Butner, S.; Cowley, W.; Gregory, M.; and Walker, J. 2009. Describing story evolution from dynamic information streams. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, 99–106.
- Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. *Automatic Keyword Extraction from Individual Documents*. John Wiley & Sons, Ltd. 1–20.
- Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851–860. ACM.
- Wigand, F. D. L. 2010. Twitter takes wing in government: diffusion, roles, and management. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, dg.o ’10, 66–71. Digital Government Society of North America.