Ersch. in: Computer / IEEE ; 46 (2013), 5. - S. 68-75

http://dx.doi.org/10.1109/MC.2012.430

Visual Analysis of Social Media Data

Tobias Schreck and Daniel Keim
University of Konstanz, Germany

The application of visual analytics, which combines the advantages of computational knowledge discovery and interactive visualization, to social media data highlights the many benefits of this integrated approach.

Recent decades have seen tremendous progress in information technology. Long-term growth rates of 30 percent and more in storage capacity, network bandwidth, and CPU processing power have made it possible to collect, store, transmit, and process unprecedented amounts of data. This has led to novel commercial, scientific, and government applications that analyze "big data" to gain new insights and improve decision making. Examples include customer relationship, financial market, demographic, and simulation data analysis.

In the social media domain, users generate content in various forms—including video, images, text, and geospatial data—that is often freely available. This data can be used for many purposes—for example, by corporations to improve business processes, by policymakers to identify trends in public opinion, and by public health officials to monitor infectious disease outbreaks or coordinate rescue efforts after a natural disaster. Social and political scientists, among other researchers, also study social media as a cultural mirror.

Leveraging social media data, however, presents many challenges. It is large in volume and generally transmitted in high-frequency streams. The data is also multimodal, often ambiguous in its content, and highly context- and user-dependent. Communication patterns change rapidly within and among the various types of social media.

Established computer science disciplines provide answers to some of the problems arising in processing and making sense of such large, complex data. Computational knowledge discovery uses machine learning methods to automatically detect those patterns in data that can be specified algorithmically. Interactive visualization maps complex data to visual forms that users can navigate to discover interesting relationships.

The emerging field of *visual analytics* combines the advantages of both approaches to better understand big data: knowledge discovery reduces the size of the data to be inspected by filtering out the most interesting structures, while interactive visualization stimulates analysts' creativity and exploits their background knowledge. The "Elements of Visual Analytics Systems" sidebar discusses key methodological properties of these systems.

The application of visual analytics to social media data highlights the many benefits of this integrated approach.

SOCIAL MEDIA DATA

We define social media to include all media formats by which groups of users interact to produce, share, and augment information in a distributed, networked, and parallel process. The most popular examples include Twitter (short text messages), blogs and discussion forums (commentary and discourse), Flickr (photos), YouTube (videos), and OpenStreetMap (geospatial data). A common feature of these services is that users can form interest groups or other types of connections (such as leader/follower in Twitter), giving rise to relationship properties in addition to content. Media such as Facebook, XING, and LinkedIn specifically serve to form social networks.

Social media produces tremendous amounts of data that can contain valuable information in many contexts. Moreover, anyone can access this data either freely or by means of subscriptions or provided service interfaces, enabling completely new applications. For example, researchers could use such data to track opinions about new products and services, fads and trends in popular culture, adverse reactions to prescription drugs, infectious disease epidemiology, fraud and other types of criminal activity, the public's response to a political candidate or proposed legislation, motor vehicle defects, and different groups' consumption habits.

However, processing many social data formats is complex. Twitter messages, for example, are very short, making analysis of message context difficult. In addition, social media often evolve dynamically and produce new forms of language—jargon, abbreviations, and so on—that are difficult to process using existing text analysis methods.[1] The semantic gap—the use of different linguistic descriptions of an object—likewise makes it difficult to assess content. Moreover, data types are often mixed, compounding complexity.[2] For example, many Twitter messages and forum postings contain hyperlinks to Web media or geo- and time-stamp information. Finally, the large volume of data generated by social media necessitates efficient and stream-oriented processing approaches.

In general, these problems can make fully automatic analysis of social media data impractical. The analysis often requires exploratory approaches, and user involvement is crucial. To that end, visual representations can help provide an initial overview of the data, enabling an analyst to identify and navigate particular aspects of interest. Visual feedback on the impact and dependency of various analysis parameters can also foster effective analysis—for example, by resolving ambiguities in text or media recognition or setting the scale of the analysis in space and time.

VISUALIZING SOCIAL MEDIA DATA

The proliferation of social media data and the opportunities it offers to various stakeholders have sparked considerable work in visual analytics. Researchers have prototyped numerous systems that extract multifaceted information from social media data streams; correlate this information with textual, geospatial, temporal, and other contextual data; and present the results in novel,

Elements of Visual Analytics Systems

Systems for visual analysis of large, complex data integrate computational knowledge discovery with interactive visualization.

Computational knowledge discovery uses efficient data mining methods to generate patterns from large datasets.[1] Examples include grouping data into clusters or predicting class labels for unseen data. Visualization involves generating cognitively useful visual representations of data, such as diagrams and graphs. The goal is to use the human visual sense as a large-bandwidth channel to perceive information and stimulate creativity in thinking about its meaning.[2] In scientific data, 2D/3D coordinates often form the basis of visual designs,[3] while in social media data, network graphs often encode relationships between news items or users in a community.[4,5]

Visual analytics also draws on cognitive psychology and perception research to guide the design of workflows, human-computer interaction research to provide effective design and evaluation criteria for user interfaces, and database management systems research to help deal with large amounts of data efficiently and reliably.

Solving real-world problems often requires combining visual analytics methods and adapting them to a specific domain or use case.[6,7] For example, exploratory data analysis employs visual techniques to deal with problems in statistics,[8] while visual data mining uses visualization during data preprocessing or to present analytical results.[9]

References

1. U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad et al., eds., AAAI Press, 1996, pp. 1-34.
2. S.K. Card, J.D. Mackinlay, and B. Shneiderman, eds., *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999.
3. C.D. Hansen and C.R. Johnson, eds., *Visualization Handbook*, Academic Press, 2004.
4. U. Brandes and T. Erlebach, eds., *Network Analysis: Methodological Foundations*, LNCS 3418, Springer, 2005.
5. V. Geroimenko and C. Chen, eds., *Visualizing the Semantic Web: XML-Based Internet and Information Visualization*, 2nd ed., Springer, 2006.
6. J.J. Thomas and K.A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, Nat'l Visualization and Analytics Center, 2005.
7. D.A. Keim et al., eds., *Mastering the Information Age: Solving Problems with Visual Analytics*, Eurographics Assoc., 2010.
8. J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.
9. S.J. Simoff, M.H. Böhlen, and A. Mazeika, eds., *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, LNCS 4404, Springer, 2008.

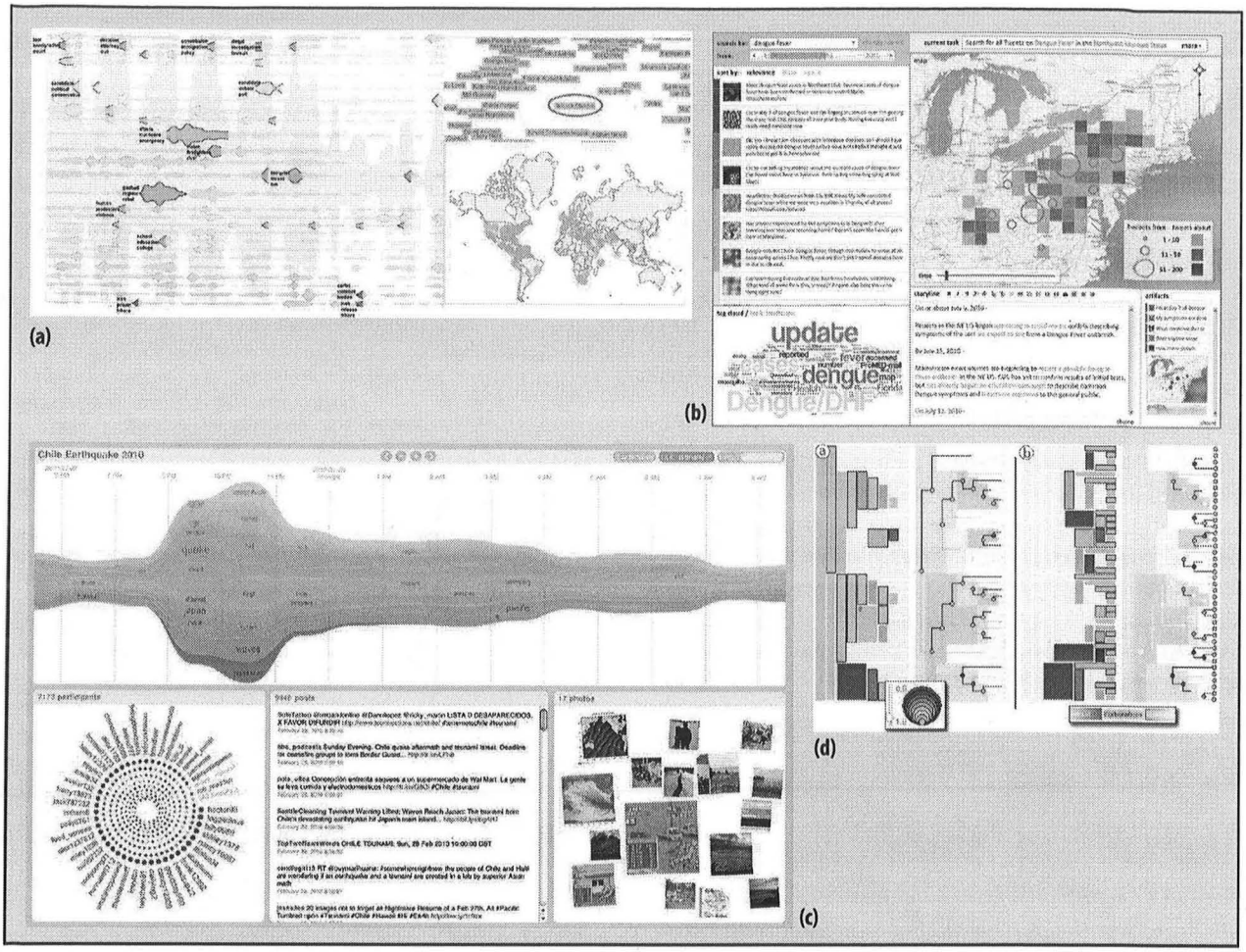

Figure 1. Prototype visual analytics systems for social media data. (a) LeadLine applies text-processing techniques to extract events from social media data streams and characterize them. (b) SensePlace2 provides situational awareness support for emergency response in the form of a visual search and monitoring tool for geolocated Twitter data about events such as natural disasters and pandemics. (c) A proposed "visual backchannel" integrates text, attached images, and author identities from Twitter data. (d) Node-link trees of rhetorical structures can help researchers parse online discourse in social media. (Images reprinted with permission.)

interactive visual displays. Figure 1 illustrates several examples.

LeadLine, shown in Figure 1a, applies text-processing techniques to extract events from social media data streams and characterize these events using combinations of topic, person, location, and point in time—in this case, events related to President Obama. It then displays the results on a graph, with related events highlighted by color-coded bursts.[3]

Based on a requirements analysis by the emergency-response community, SensePlace2, shown in Figure 1b, provides situational awareness support in the form of a visual search and monitoring tool for geolocated Twitter data about events such as natural disasters and pandemics—for example, a supposed Dengue fever outbreak in the northeastern US.[4]

Figure 1c shows a proposed "visual backchannel" that integrates text, attached images, and author identities from Twitter data. The system enables the visual exploration and monitoring of evolving online discussions about major events such as political speeches, natural disasters, and sporting competitions using Topic Streams (top), a temporally adjusted stacked graph of topics; a People Spiral indicating participants and their activities (bottom left); a traditional post listing (bottom center); and an Image Cloud of popular photos arranged by size (bottom right).[5]

Social media content is often produced as part of an ongoing communication process between the members of a community. Node-link trees of rhetorical structures, like those in Figure 1d, can help researchers parse such discourse.[6]

In our own work, we have explored new and interesting ways to visually analyze many types of social media data including community-provided photo collections, streaming news data, and georeferenced microblog data.

COMMUNITY-PROVIDED PHOTO COLLECTIONS

Social media sites such as Flickr and Panoramio enable users around the world to share images, leading to the formation of large photo repositories. Flickr, for example, announced that it exceeded six billion uploaded photos in August 2011, and it currently adds more than 1.4 million public images per day (www.flickr.com/photos/franckmichel/6855169886).

Such images often include time and location information, either as user annotations or embedded temporal and GPS metadata, that can reveal patterns about the distribution and relation of photos, users, and places across time. Visual analytics makes it possible to set parameters for exploring such data—for example, to survey an area for the most popular places visited by tourists, assuming the number of photos taken is an indicator of popularity.

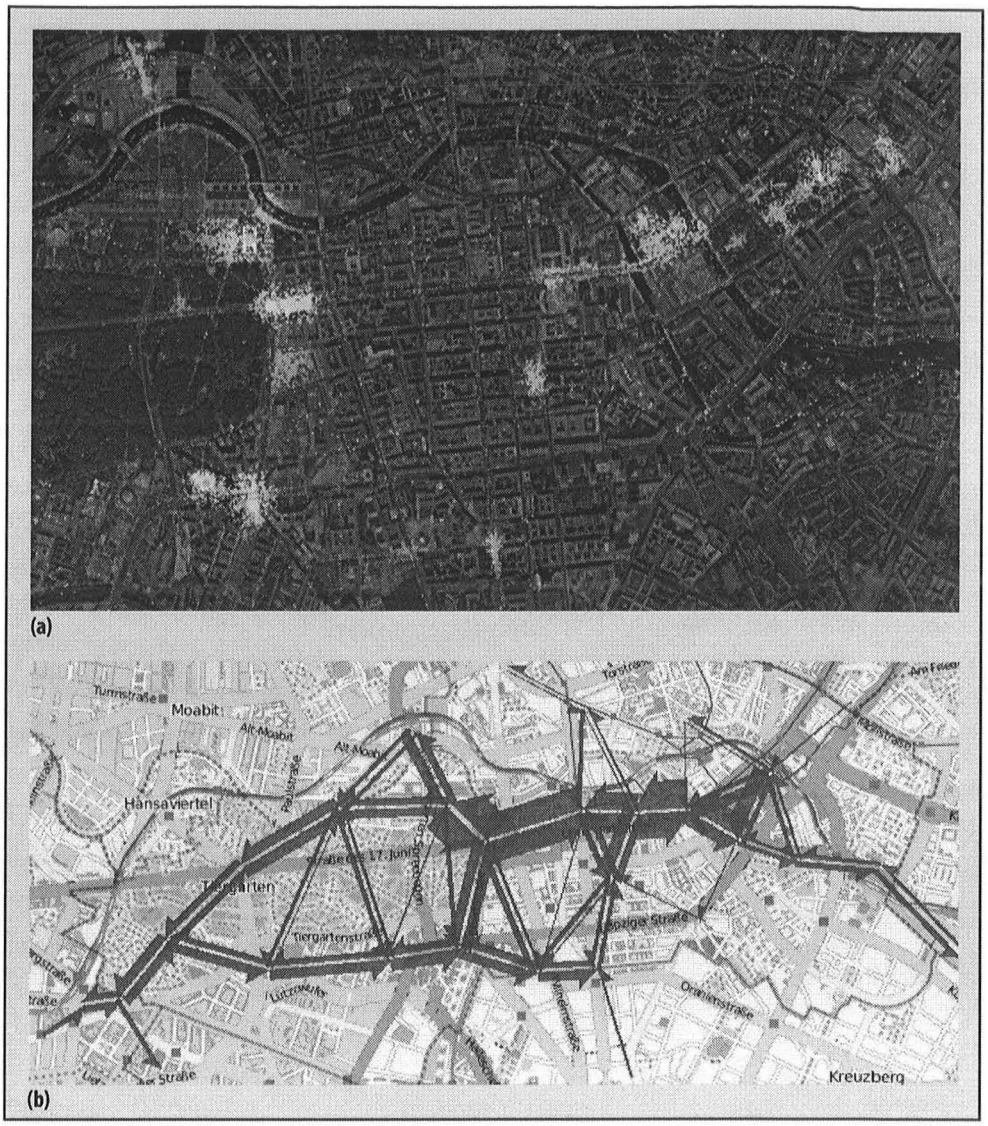

Figure 2. Applying visual analytics to community-provided photos of central Berlin. (a) Heat map of the distribution of images, obtained by kernel density estimation. Red and white areas denote places with a high density of images—that is, the most popular locations. (b) Flow map of paths followed by photographers, computed by means of trajectory cluster analysis. Arrows denote flow direction and magnitude, revealing a bidirectional northern flow and a directional southern flow. (Source: G. Andrienko et al., "Analysis of Community-Contributed Space- and Time-Referenced Data by Example of Panoramio Photos," *Proc. Vision, Modeling, and Visualization Workshop* [VMV 09], Inst. für Simulation und Graphik, 2009; http://vmv09.tu-bs.de/downloads/papers/and09.pdf.)

Researchers from the University of Bonn and the University of Konstanz studied a set of nearly 600,000 Panoramio photos taken in Germany between 2005 and 2009.[7] Figure 2a shows a heat map of the distribution of images in central Berlin, obtained by kernel density estimation. Red and white areas denote places with a high density of images, including popular attractions such as the Brandenburg Gate and the Reichstag building. This simple visual analysis would be useful for tourism and city planning purposes, among others.

Using photo time stamps, it is also possible to extract the paths individual photographers follow and, from the set of all paths, compute the most salient flows by means of trajectory cluster analysis. Figure 2b shows a flow map for the Berlin area, in which the most popular routes between points of interest are easily recognizable. Interestingly, the main northern flow is bidirectional, while the southern flow largely goes from west to east. Again, such information would be valuable for scenarios such as evaluating traffic congestion or scheduling road improvements.

The results presented here exclusively relied on provided image metadata. More advanced image recognition

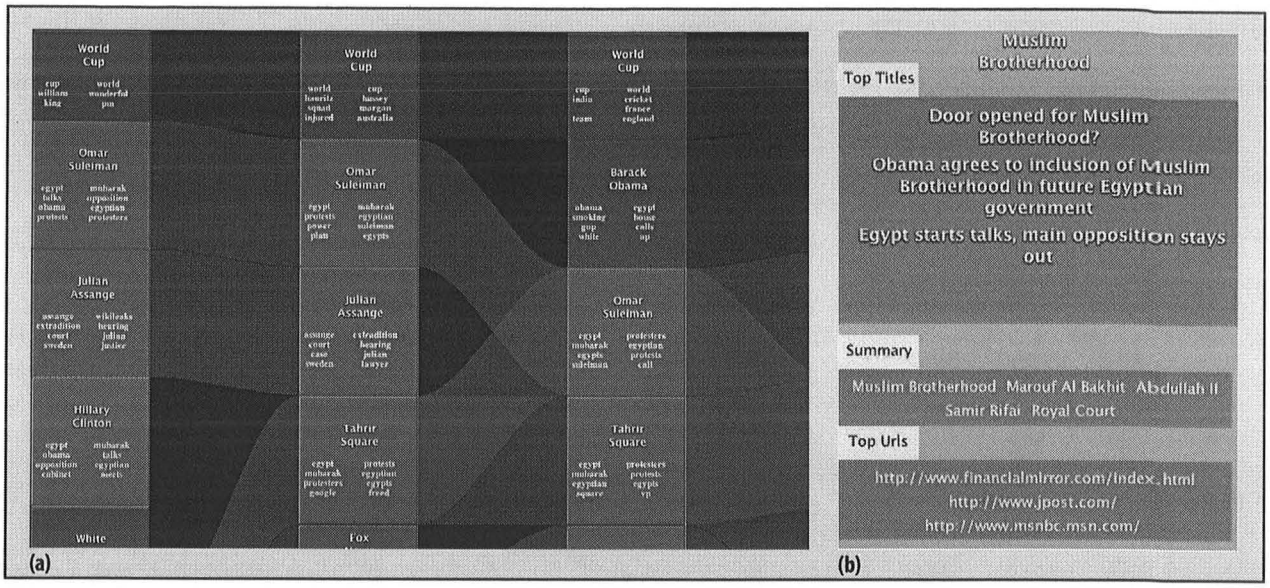

Figure 3. Node-link-based overview of a stream of news. (a) Keywords highlight the most important topics in columnar form for each day. Transparent connectors indicate entering, exiting, merging, and splitting topics over time. (b) Zoom views reveal increasingly more detail of news coverage. (Source: M. Krstajić et al., "Incremental Visual Text Analytics of News Story Development," *Proc. SPIE*, vol. 8294, 2012; doi:10.1117/12.912456).

functions could be employed to enhance analysis of the photos—for example, to determine weather conditions, the presence of groups of persons or vehicles, and other visible properties of interest.

STREAMING NEWS DATA

Much valuable online information is given in textual form. Worldwide, newspapers, magazines, news agencies, and other media outlets publish tens of thousands of news articles every day. These articles include material of common interest, such as political events, as well as specialized interest, such as business reports on specific corporations. Various institutional and personal blog services also generate less formal news streams.

The amount of published textual content on the Web is enormous and ever expanding, making it increasingly difficult for analysts to stay on top of stories and events. Fortunately, text processing and mining research has made large strides in recent years—for example, new methods are available to compare documents for similar content, extract named entities relating to persons and places, and identify common topics among sets of documents. Combining these techniques with interactive visualization enables the tracking of subjects in streams of text over time.

Researchers at the University of Konstanz have developed one such system, which first groups documents in streaming news data based on similarity with respect to textual content and then extracts the most important topics.[8] The system displays the results in columnar form, with keywords indicating the main content of news in a given day.

Over time, the distribution of topics in news stories or blogs changes: a given topic can receive more coverage, seed subtopics, or fade out of view, while new topics emerge. The system detects this topical evolution and, using a node-link-based visualization, can show the flow of entering, exiting, merging, and splitting topics.

Figure 3a shows a flow view of the most important topics extracted from a set of news articles from four days in early 2011. At the center of attention was a protest movement against then-President Mubarak in Egypt. As the two leftmost columns show, news during the first two days involved discussion of a possible new president, Omar Suleiman, and reports of protests in Cairo's Tahrir Square. During the third day, news coverage of these events merged, indicating that the stories became more closely related over time. Other major stories taking place in parallel included the World Cup competition and an extradition hearing in London for WikiLeaks founder Julian Assange.

The flow view can only capture the most discriminative keywords as labels. As Figure 3b shows, users can zoom in on news topics for more details—including article headlines, article summaries, and top-ranked URLs of article sources.

GEOREFERENCED MICROBLOG DATA

Microblogging services such as Twitter stream millions of user messages every day. With the advent of nearly ubiquitous mobile Internet access in many countries and the availability of low-cost GPS sensors for end users, these messages are expected to increasingly carry geotagged information that researchers can analyze along with textual content.

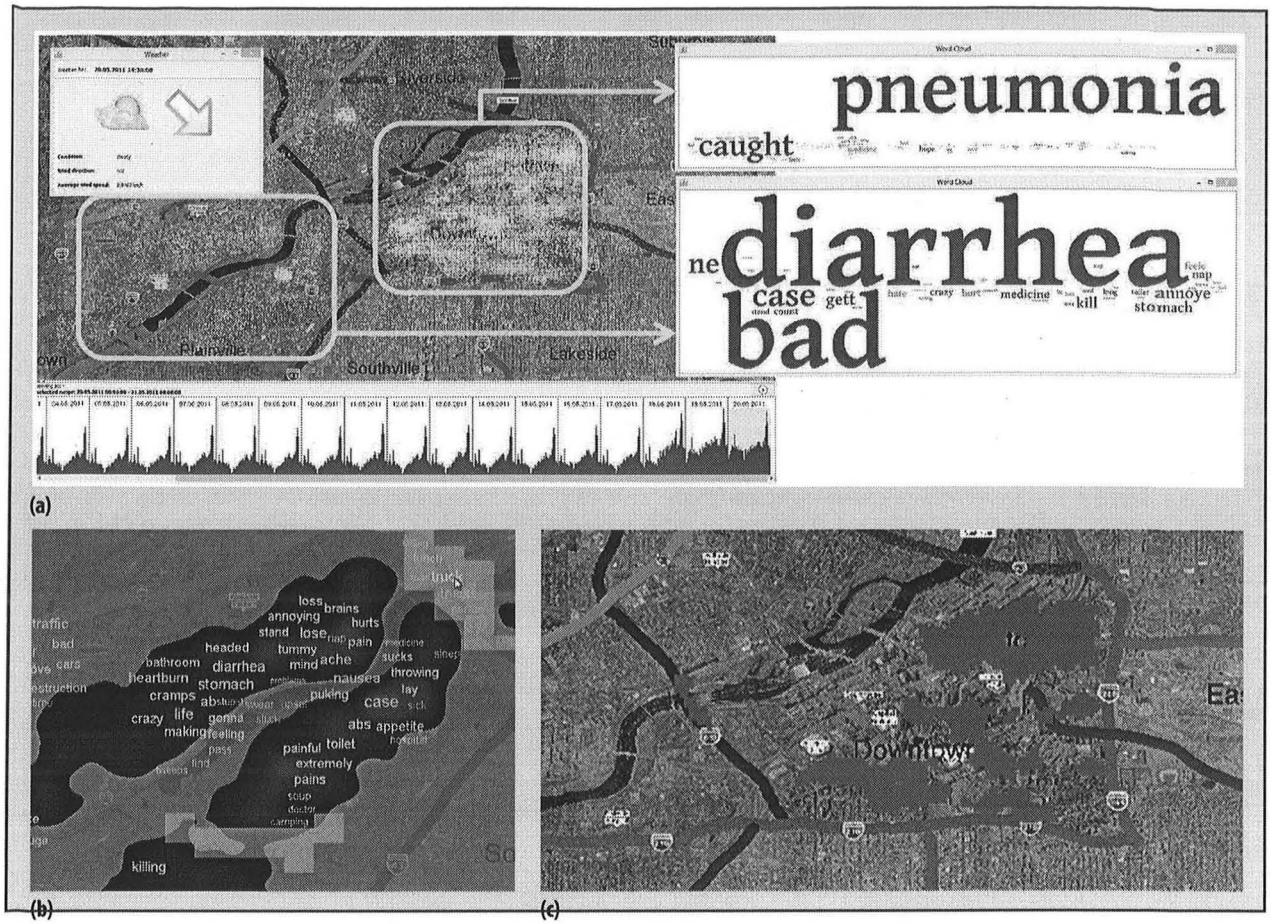

Figure 4. Using visual analytics to explore geotagged microblog data about a fictitious epidemic outbreak. (a) Textual overviews in the form of word clouds, together with interactive filtering, reveal interesting patterns. (Source: E. Bertini et al., "Visual Analytics of Terrorist Activities Related to Epidemics," *Proc. IEEE Conf. Visual Analytics Science and Technology* [VAST 11], IEEE, 2011, pp. 329-330.) (b) ScatterBlog visualization integrates word clouds with a traditional map. (Source: H. Bosch et al., "ScatterBlogs: Geo-spatial Document Analysis," *Proc. IEEE Conf. Visual Analytics Science and Technology* [VAST 11], IEEE, 2011, pp. 309-310.) (c) Automatic anomaly detection compares message content and frequency with historic data to highlight potentially important time and space patterns.

In 2011, in one of three minichallenges conducted as part of the IEEE VAST Challenge, held in conjunction with the Conference on Visual Analytics Science and Technology, participating research teams were charged with analyzing about one million geolocated Twitter messages to characterize a flu-like epidemic outbreak in the fictitious city of Vastopolis (www.cs.umd.edu/hcil/vastchallenge). A team from the University of Konstanz developed an interactive display that filters data by location and time. By analyzing the daily occurrence of messages, the researchers could trace the epidemic's progress based on changes in message volume at city hospitals.[9]

Figure 4a shows the main interface, which displays messages in a dot map. Investigation of message content using word clouds revealed important keywords that described symptoms of the disease, such as diarrhea and pneumonia. In response to the same minichallenge, Stuttgart University researchers developed the ScatterBlog technique, shown in Figure 4b, which directly embeds the most discriminative keywords in a traditional map display.[10]

The University of Konstanz system also includes automatic analysis components. As Figure 4c shows, an anomaly detection method compares message content and frequency with historic data to highlight potentially important time and space patterns. In general, user feedback is needed to define the phenomena of interest. Automatic analysis can successfully detect anomalies, but users must explore and refine the underlying reasons.

TOWARD THE FUTURE

Visual analytics is a frontier research topic. Combining interactive visualization with automatic data analysis affords new and exciting capabilities, and many domains would benefit from the application of this methodology to large, complex data like that generated by social media.

Visual Analytics Research Forums

Numerous forums advance the science of visual analytics for use on social media data and many other types of information. Challenges and evaluation campaigns also motivate researchers to develop, test, and demonstrate new methods.

The largest visualization-oriented forum is IEEE VIS (http://ieeevis.org), which brings together researchers in academia, government, and industry for three conferences: the Conference on Visual Analytics Science and Technology (VAST), the Information and Visualization Conference (InfoVis), and the Scientific Visualization Conference (SciVis). Formerly known as VisWeek, this weeklong gathering includes tutorials, workshops, panels, and a doctoral colloquium.

Visualization-related workshops and symposia are colocated at IEEE VIS; this year's gathering in Atlanta, Georgia, in October will feature the Symposium on Large-Scale Data Analysis and Visualization (LDAV), the Symposium on Biological Data Visualization (BioVis), and the Workshop on Visualization for Cyber Security (VizSec). Previous events concentrated on analysis of social media data, including the IEEE VisWeek workshops on Interactive Visual Text Analytics for Decision Making (2011) and on Task-Driven Analysis of Social Media Content (2012).

Other conferences contribute to work on visual analytics including SIGGRAPH, convened by the ACM Special Interest Group on Graphics and Interactive Techniques (www.siggraph.org), and the Eurographics Conference on Visualization (www.eurovis2013.de). In addition, various research labs and organizations do work on visual analytics. Some of these focus on social media, such as the International Network for Social Network Analysis (www.insna.org).

Prototype visual analytics systems and techniques can be evaluated using various benchmark datasets. In the case of social media, the 2011 IEEE VAST Challenge provided a large-scale Twitter and text news corpus for researchers to track a fictitious epidemic outbreak (www.cs.umd.edu/hcil/vastchallenge). Similarly, the RepLab 2012 evaluation campaign called on researchers to score the reputation of corporations based on opinions expressed in a carefully compiled Twitter dataset (www.limosine-project.eu/events/replab2012).

At the same time, researchers must overcome several challenges.

Visual analytics system designers must work to strike a better tradeoff between automatic analysis and manual interactive exploration. Today's systems often focus too much on one of these aspects. However, in many cases this is not the best distribution of work between the computer, which excels at automatic analysis, and human analysts, whose creativity and visual recognition capabilities are unmatched by machines. To solve real-world problems, users and computers must cooperate in more harmoniously integrated workflows that maximize the relative analytical capabilities of humans and machines. For example, the system could automatically determine the best analysis steps and suggest promising data views for the user to explore.

In addition, visual analytics needs a more refined taxonomy. In a simplified model, the design space for a current system can be seen as consisting of five dimensions—application requirements; data types; and visualization, data mining, and user interaction techniques—that researchers have only begun to map. For example, Ben Shneiderman proposed a data-type-by-task taxonomy to guide research in information visualization,[11] but more work in this area is needed. With the help of extended taxonomies and lessons learned from applications, visual analytics designers will better understand how to model functional dependencies between different dimensions.

Understanding complex data requires a holistic approach.[12] As the "Visual Analytics Research Forums" sidebar describes, forums bringing together diverse scientists from around the world have emerged to address the challenges of visual analytics and its application to different types of data, including social media data.

We are just on the cusp of being able to extract useful knowledge from social media data. Advances in natural language processing and information retrieval will make it possible to scale these techniques to larger data streams, parse community-specific terminology, and integrate different data types—for example, user profile relationships, embedded or linked multimedia data, and current events from news sources. At the same time, progress in information visualization will make it easier for users to monitor and explore social media data, dynamically adapting to their background knowledge as well as their operational and analytical skills.

Visual analysis of social media could enable a wide range of promising new applications. For example, such analysis could provide situational awareness information for disaster response organizations, helping them to better understand an evolving situation, allocate limited resources, and coordinate rescue actions. The international relief effort following the 2010 Haiti earthquake was the first to rely substantially on social media—aid organizations used Twitter to plan and monitor relief operations, and the OpenStreetMap community helped map out the country's affected infrastructure within days, providing invaluable support to relief organizations. Better visual analytics tools would enable both domain experts and amateur volunteers to leverage social media data in this and many other areas.

Acknowledgments

We thank Natalia and Gennady Andrienko, Milos Krstajić, Slava Kisilevich, Florian Mansmann, and Enrico Bertini for their valuable collaborations. We also thank Daniela Oelke, Christian Rohrdantz, and Andreas Karsten for many fruitful discussions.

References

1. K. Bontcheva and D. Rout, "Making Sense of Social Media Streams through Semantics: A Survey," *Semantic Web*, to appear, 2013; http://semantic-web-journal.org/sites/default/files/swj303_0.pdf.
2. T. von Landesberger et al., "Visual Search and Analysis in Complex Information Spaces: Approaches and Research Challenges," *Expanding the Frontiers of Visual Analytics and Visualization*, J. Dill et al., eds., Springer, 2012, pp. 45-67.
3. W. Dou et al., "LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration," *Proc. IEEE Conf. Visual Analytics Science and Technology* (VAST 12), IEEE, 2012, pp. 93-102.
4. A.M. MacEachren et al., "SensePlace2: GeoTwitter Analytics Support for Situational Awareness," *Proc. IEEE Conf. Visual Analytics Science and Technology* (VAST 11), IEEE, 2011, pp. 181-190.
5. M. Doerk et al., "A Visual Backchannel for Large-Scale Events," *IEEE Trans. Visualization and Computer Graphics*, vol. 16, no. 6, 2010, pp. 1129-1138.
6. J. Zhao et al., "Facilitating Discourse Analysis with Interactive Visualization," *IEEE Trans. Visualization and Computer Graphics*, vol. 18, no. 12, 2012, pp. 2639-2648.
7. G. Andrienko et al., "Analysis of Community-Contributed Space- and Time-Referenced Data by Example of Panoramio Photos," *Proc. Workshop Vision, Modeling, and Visualization* (VMV 09), Inst. für Simulation und Graphik, 2009; http://vmv09.tu-bs.de/downloads/papers/and09.pdf.
8. M. Krstajić et al., "Incremental Visual Text Analytics of News Story Development," *Proc. SPIE*, vol. 8294, 2012; doi:10.1117/12.912456.
9. E. Bertini et al., "Visual Analytics of Terrorist Activities Related to Epidemics," *Proc. IEEE Conf. Visual Analytics Science and Technology* (VAST 11), IEEE, 2011, pp. 329-330.
10. H. Bosch et al., "ScatterBlogs: Geo-Spatial Document Analysis," *Proc. IEEE Conf. Visual Analytics Science and Technology* (VAST 11), IEEE, 2011, pp. 309-310.
11. B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proc. IEEE Symp. Visual Languages* (VL 96), IEEE CS, 1996, pp. 336-343.
12. J. Kehrer and H. Hauser, "Visualization and Visual Analysis of Multi-faceted Scientific Data: A Survey," *IEEE Trans. Visualization and Computer Graphics*, to appear; www.ii.uib.no/vis/publications/publication/2012/pdfs/Kehrer12VisualizationAnd.pdf.

Tobias Schreck *is an assistant professor of visual analytics in the Department of Computer and Information Science at the University of Konstanz, Germany. His research interests include visual search and analysis of time-oriented, high-dimensional, and 3D object data, with applications in data analysis and multimedia retrieval. Schreck received a PhD in computer science from the University of Konstanz. He is a member of the IEEE Computer Society. Contact him at tobias.schreck@uni-konstanz.de.*

Daniel Keim *is a full professor and chair of the Visualization and Data Analysis group in the Department of Computer and Information Science at the University of Konstanz, Germany. His research interests include visual analytics, information visualization, and data mining. Keim received a PhD in computer science from the University of Munich, Germany. He is a member of the IEEE Computer Society. Contact him at daniel.keim@uni-konstanz.de.*