EuroVis Workshop on Visual Analytics (2013), pp. 1–5
M. Pohl and H. Schumann (Editors)

Social Media Analytics for Competitive Advantage

William Ribarsky, †1 Derek Xiaoyu Wang, ‡ 1 and Wenwen Dou, § 1

[1]The Charlotte Visualization Center
UNC Charlotte, Charlotte, NC, USA

Abstract

Big Data Analytics is getting a great deal of attention in the business and government communities. If it lives up to its name, visual analytics will be a prime path by which visualization competes successfully in this arena. This paper discusses some fundamental work we have done in this area through integration of interactive visualization and automated analysis methods and the applications that have resulted.

Categories and Subject Descriptors (according to ACM CCS): Visualization [I.6.9.A]: —

1. Introduction

As we know, big data analytics has become more than a term, it is now a movement. Whether or not "big data" is a buzzword, there is reason to believe that interest in the broader issues surrounding not just scalably large data but even more so data that are complex and can be associated with problems that require complex reasoning will not abate soon and could even grow stronger. There are many reasons for this. More companies (and government agencies, too) are building comprehensive and long-term databases. But there is a growing realization that traditional database techniques, though they are in principle scalable and useful for many things, cannot tell you basic things about what the database contains and what the important relations and trends are [Oli]. (For example, we have worked with a bank that has a very large, comprehensive transactional database yet struggles to use all the value it contains.) In addition, the advent of social media and online sources show that useful data can come from anywhere, inside or outside the company. Due to the availability of all these data, there is a growing push to increase data-driven decision-making, but this is hampered, as inferred above, by not knowing what actionable information the data contains. Finally, business studies indicate that timely, effective use of data-derived knowledge is a competitive advantage and that not using this knowledge effectively is a competitive disadvantage [MCB11, LL11, Thi12]. Companies who do not marshal their data resources will be losers in the long term. In fact, those that find new uses for their or other data will be the biggest winners.

Another reason that big data analytics will have staying power is that a robust infrastructure is being built. Based on surveys in the U.S. alone, McKinsey estimates that there will be a deficit of nearly 200K professionals with deep analytics skills by 2018 and the need to retrain 1.4M managers so that they understand the value of data and know the right questions to ask [MCB11]. Gartner estimates that 1.9M big data jobs will be created in the U.S. by 2015 [Thi12]. It will be hard to fill even a fraction of this need at the current rate of production for data analytics professionals.

The premise of this paper is that visualization and especially visual analytics is ideally positioned to take advantage of this opportunity. By its nature, visual analytics supports exploration, discovery, and complex reasoning about data and data-driven problems. Statistical, data mining, machine learning, signal processing, and other deep analytics methods are tightly integrated with interactive visualizations. Visual analytics aims to put the human in the loop at just the right point to discover key insights, develop deep understanding, make decisions, and take effective action. On both the visualization side and the analytics side, visual analytics is positioned to effectively manage and support understanding of scalably large data, and this promise is being made concrete as new techniques are developed [WDM*12]. Because of the central role of visual analytics, there is also the opportunity to imbue the massive influx of new data analyt-

† Email to: ribarsky@uncc.edu
‡ Email to: xiaoyu.wang@uncc.edu
§ Email to: wenwen.dou@uncc.edu

ics professionals with knowledge of visual analytics through courses and training. This should definitely be a key part of the visual analytics academic agenda; it will result in a whole generation of analysts, engineers, and managers who appreciate and know how to use these tools.

In this paper, we will address how competitive advantage can be derived from the analysis of unstructured data, especially social media data (though the techniques used can be applied to a broader range of unstructured data). We will use mostly examples from our own work, though there is a range of other work.

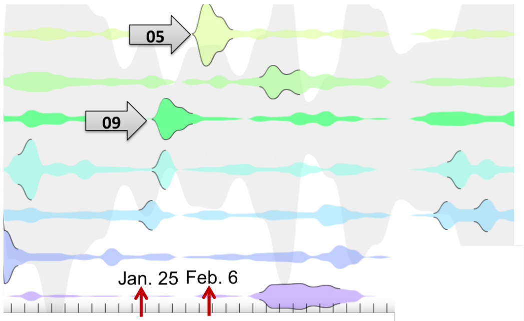
2. The Nature of the Data

We focus on streaming Twitter data in this paper. We have been collecting a 1% random sample of these data for nearly 1.75 years. This results in a large number of tweets (now in excess of 20B). We have used this Twitter collection as a testbed for several recent studies with diverse subjects [WDM*12, DWS*12, WDMR13, DWL*13] including the ones reported here. We are now building a collection of texts from Facebook posts, which will permit us to explore a different demographic range than that for Twitter.

Text messages from Twitter, Facebook, and several other social media services have general attributes such as unstructured content and intrinsic uncertainty as to the validity of the messages. In addition, these data have the attributes of data physicality and data sociality. The messages are often intimately connected with particular times and locations (either locating where and when the message was sent or by mention of places and dates, either past, present, or future in the message body. Of course, social media messages are sent, received, or re-sent by people, so there can be a rich social connectivity revealed. The availability of such information, often minute-by-minute or over the whole length of a story that may take months to unfold, is a new and very powerful aspect of social media analyses.

3. Topic Modeling and Entity Extraction

To provide meaning and organization to the unstructured data, we use Latent Dirichlet Allocation (LDA) [BNJ03, DWCR11], which reveals latent topics from large text collections, which are then described by coherent sets of keywords (with the leading words being the most meaningful for the topic, as illustrated in Figure 1). To this we add named entity recognition, based on a customized dictionary and the use of LingPipe with statistical chunking. This permits the identification of people, locations, buildings, times, dates, etc. from within the text messages. We have extended the traditional LDA approach to handle temporal features and structure (in particular events, as described further below). We have also developed scalable capabilities for efficiently generating topics even for very large text collections [WDM*12]. We have successfully applied these techniques to a range of text collections including project abstracts, reports, research papers, streaming Twitter data, and recently patent descriptions. This set of approaches provides the ability to attack unstructured data both inside and outside the company, conferring competitive advantage [NN12].



Topic 09 heart launch fragrance donate fight@selenagomez: @goredfor women…
Topic 05 nmbeautyevent beauty@belk parfum summer cream..

Figure 1: *Event view of Macy's women's heart donation campaign (Topic 09) and Belk/ Neiman Marcus beauty products promotion (Topic 05). Each horizontal ribbon indicates a topic stream, with time runs from left to right. Each ribbon is corresponded with a topic listed below. The black contour outlines events within each topic stream.*

4. Events and Time Structuring

The fundamental component of our time structuring is the event, which we define as a "meaningful occurrence in space and time". Events are bursts of activity over a relatively short time period, the time scale depending on the category of the temporal data. For example, with streaming Twitter data, a typical single event burst lasts one to two days; major events can be longer lasting, but they usually can be divided into sub-events. In this paper events are associated with a particular topic (as shown in Figure 2) so that an event occurs for a particular topic, time, and set of extracted entities (e.g., location, indicated past or future times, names of people, etc. extracted from the social media texts). Thus in the case of the interactive interface we have developed for Twitter data, a selection of an event chooses only those tweets for the given topic and for the part of the event burst time range selected. As discussed below, events provide a great focus and together make up an interpretable narrative; thus this selection is a powerful filtering tool.

We perform one more analysis step on our event structure. We label as events only those bursty structures that have a motivating event (see Figure 2). A motivating event is an occurrence, either described in the event burst tweets themselves (usually at the beginning) or external to this set of tweets, that has motivated the bursty response. Most if not all event bursts of this type are responses to the initial

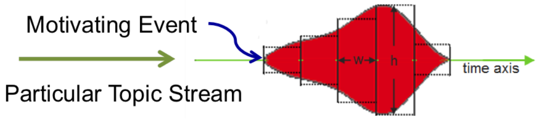


Figure 2: *Bursty structure for an event. H indicates the volume of documents (tweets) associated in this event, and W suggests the duration of the event.*

motivating event. For example, the main topical events on September 17, 2011 were clearly associated with the launch of Occupy Wall Street (OWS) on that date at Zucotti Park in New York City, but most of the associated tweets, from individuals and from online news, were in response to this event. In fact, OWS was large enough that there were several topics with their associated events on that date. We have found that by just analyzing the shape, size, and duration of the burst, we can automatically identify events that will have clear motivating events [LYK*12]. (These are the bursts with dark outlines in Figure 1 and subsequent figures.) Thus we have a mechanism for automatically identifying meaningful events that we have tested successfully on multiple categories of data, not just streaming social media. This is not to say that there are not other, unmarked bursts that are meaningful. Nor is it to say that the meaning is immediately clear from this analysis. Input from a human-in-the-loop is necessary to resolve these questions. But this identification of meaningful events is still a boon for exploratory analysis since we have found it identifies most of the major events and also directs the user's attention. We have applied all the techniques in Secs. 3 and 4 to tell the complete story of OWS from precursor discussion before the launch till now. This shows how a comprehensive, rich narrative can be built efficiently [WDMR13].

5. Deriving Competitive Advantage

In the rest of this paper, we discuss some business cases that show how competitive advantage might be derived. We first did a study of department stores in the Charlotte region (Belk, Macy's, Dillard's, Neiman Marcus, and Saks Fifth Avenue) by using messages with hash tags or entities naming the stores and then bringing in related tweets through topic modeling. The study was over a 10 month period starting in Fall, 2011. About 15 main topics resulted for this time period. They immediately revealed some general information. Twitter response often had to do with marketing around celebrations, charitable campaigns, and holiday events (e.g., Macy's Thanksgivings Day parade). There were often tie-ins to women's cosmetics and beauty products. Macy's had the largest Twitter presence during the period followed by Neiman Marcus. Belk had a growing presence later in the period. Dillard's never had much of a presence.

Figures 1 and 3 show specific events illustrating how detailed analyses over time can be done (in this example restricted to a 3 month period at the beginning of 2012). The presented events plus other events that fill in the Twitter story for these stores were generated automatically and then quickly studied in more detail by selecting associated tweets for each topic and event. Figure 1 shows response to a marketing effort associated with the go red for women heart campaign sponsored by Macy's (Topic 09). It also shows merged responses to two similar marketing efforts launched by Neiman Marcus and Belk on beauty products that started during Super Bowl week and then continued for another week or two. In both cases, the events can easily be identified by looking at the lead words in the topic lists and some of the associated tweets.

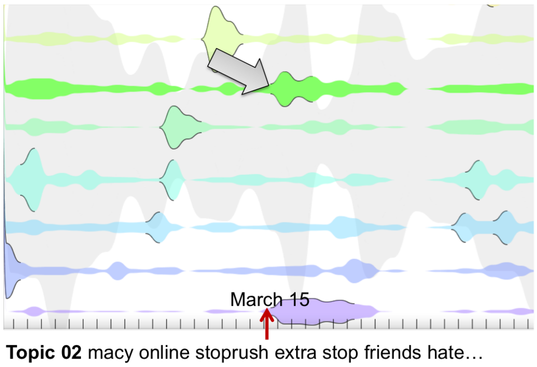
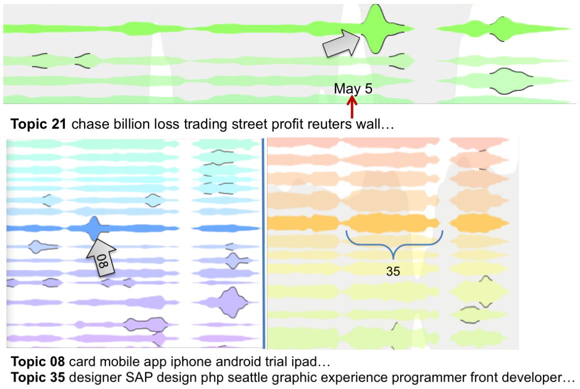


Figure 3: *Burst of negative responses to Rush Limbaugh's comments (associated with Macy online). Grey arrow point at the burst #stoprush campaign.*

Figure 3 shows another type of event with a different temporal structure. The event is conservative talk show host Rush Limbaugh's diatribe against a Georgetown University student because of her stand in favor of access to birth control, and the response to his comments. The motivating event caused a firestorm of comments against Limbaugh including a burst of activity having to do with the StopRush Twitter movement and its campaign to boycott sponsors of Limbaugh's show. Macy's became embroiled because of its sponsorship. Although the event has to do with the specific outburst about the student, other bursts and events in the same topic stream show other public complaints against Limbaugh both before and after the event in Figure 4. This stream of related activity goes over a period of months and shows that negative events and associations, even if inadvertent, can have a long effect.

Figures 4 (A) and 4 (B) bottom show details from a set of topical analyses having to do with several banks, including those headquartered in the Charlotte region (such as Bank of America), over a period of several months. The topic modeling analysis was set up in a similar fashion as the department store analysis; the set of topics is larger. In Figure 4 (A)



Topic 21 chase billion loss trading street profit reuters wall…

Topic 08 card mobile app iphone android trial ipad…
Topic 35 designer SAP design php seattle graphic experience programmer front developer…

Figure 4: *Top: (A) Bursty event at the revelation of JP Morgan Chase multibillion dollar trading loss (the London Whale). Bottom: (B) Event revealing bank's strategies on hiring additional workforce on mobile development.*

the indicated event was motivated by the initial exposure of the multibillion dollar trading loss at JP Morgan Chase. The sharp burst begins at the day of this disclosure; there was also an associated stock market drop on that day. In addition, there is a series of events for the same topic before and especially after the disclosure that tell the unfolding story of this scandal.

Although most of the events in the banking analysis have to do with disclosures like this, ongoing fallout from the mortgage and banking collapse of 2008, and response to proposed government regulations, there are still events and topics that discuss other aspects. Figure 4 (B) gives a couple of examples. Topic 08 has to do with the development of new mobile debit and credit card apps. The event indicated is response to actions by Walmart and Target, among others, and reported in the Wall Street Journal and Forbes, to develop their own mobile payment systems. This would eliminate the middle men (the banks) and would be of great interest and possible risk to them. These analyses show that even in this case where the events are dominated by negative news about banks, the exploratory visual analytics tools still find events that tell banks about their competitive environment.

6. Conclusions

We have presented the work described here plus additional analyses to a set of business partners from retail and banking. This has generated considerable interest and feedback. The ability to analyze competitor strategies is considered important. There is a desire to know about the demographics of the people generating messages for selected events and also how retweeting spreads a message. (We are working on both these things.) There is a desire to do targeted marketing based on real-time streaming tweet analysis, and we have developed a capability in this area. More generally, companies see the opportunity to analyze the response to marketing and advertising campaigns as they unfold and also investigate what affects the public view of their brand image (which can be affected by external circumstances, as we have seen). Banks want to do emerging risk analysis using both internal and external sources. We also expect that the set of methods described in this paper will be applied in the future to internal company data.

7. Acknowledgement

This work is supported in part by the Army Research Office under contract number W911NF-13-1-0083.

References

[BNJ03] Blei D. M., Ng A. Y., Jordan M. I.: Latent dirichlet allocation. *J. Mach. Learn. Res. 3* (January 2003), 993–1022. URL: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993, doi:http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993. 2

[DWCR11] Dou W., Wang X., Chang R., Ribarsky W.: Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (oct. 2011), pp. 231 –240. doi:10.1109/VAST.2011.6102461. 2

[DWL*13] Dou W., Wang X., Li Y., Ma Z., Ribarsky W.: Hierarchical topics: Visually exploring large text collections using topic hierarchies. *IEEE Visual Analytics Science and Technology (VAST), 2013* (2013). 2

[DWS*12] Dou W., Wang X., Skau D., Ribarsky W., Zhou M. X.: Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (oct. 2012), pp. 231–240. doi:10.1109/VAST.2011.6102461. 2

[LL11] LaValle S., Lesser E.: Big data, analytics and the path from insights to value. In *MIT Sloan Managment Review* (May (2) 2011), vol. 52, pp. 21–31. 1

[LYK*12] Luo D., Yang J., Krstajic M., Ribarsky W., Keim D.: Eventriver: Visually exploring text collections with temporal references. *Visualization and Computer Graphics, IEEE Transactions on 18*, 1 (jan. 2012), 93 –105. doi:10.1109/TVCG.2010.225. 3

[MCB11] Manyika J., Chui M., Brown B.: Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institue, May 2011. 1

[NN12] Nair R., Narayanan A.: *Benefitting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage*. Booz and Company, 2012. 2

[Oli] Oliver A.: Big data woes: Which database should i use? Online. URL: InfoWorld. 1

[Thi12] Thibodeau P.: Big data to create 1.9m it jobs in u.s. by 2015. ComputerWorld, August 2012. 1

[WDM*12] Wang X., Dou W., Ma Z., Villalobos J., Chen Y., Kraft T., Ribarsky W.: I-SI: Scalable Architecture of Analyzing Latent Topical-Level Information From Social Media Data. *Computer Graphics Forum 31*, 3 (2012), 1275–1284. URL: http://diglib.eg.org/EG/CGF/volume31/issue3/v31i3pp1275-1284.pdf, doi:10.1111/j.1467-8659.2012.03120.x. 1, 2

[WDMR13] WANG X., DOU W., MA Z., RIBARSKY W.: Discover diamonds-in-the-rough using interactive visual analytics system: Tweets as a collective diary of the occupy movement. In *ICSWM 2013* (2013). 2, 3