Visualizing Co-Retweeting Behavior for Recommending Relevant Real-Time Content

[Extended Abstract]

Samantha Finn
Computer Science Department
Wellesley College
sfinn@wellesley.edu

Eni Mustafaraj
Computer Science Department
Wellesley College
emustafa@wellesley.edu

ABSTRACT

Twitter is a popular medium for discussing unfolding events in real-time. Due to the large volume of user generated data during these events, it's important to be able recommend the best content while it's fresh. Current recommendation algorithms for Twitter take into account the user's tweets and her social network, but since real-time events might be unique or unexpected, the history of a user may not be sufficient for finding the most relevant content. Additionally, for users who want to join the conversation at that specific moment (or follow it without having to create an account), the system will be faced with the cold-start problem. We propose a simple visualization technique that considers the activity of the whole community participating in the real-time discussion, by capturing their co-retweeting behavior. Such a technique depicts the big picture, allowing a user to choose content from parts of the community that share her opinions or beliefs.

1. INTRODUCTION

Presidential debates in the United States are very important events. Their TV audience ranks consistently among the highest of the year (second only to the Super Bowl). During the 2012 presidential race between President Barack Obama and challenger Mitt Romney, the three debates on October 4th, 16th and 22nd drew respectively: 67 million, 65.5 million, and 59.2 million spectators. But these spectators are no longer passive. They increasingly use the web as a platform for further engagement. As studies from Pew Research have shown, 1 in 10 spectators in such debates is a dual-screener [3]. Very often, the second screen is Twitter, where running commentary of live-televised events is at its liveliest. But such lively online discussions have a big drawback: their size. The three 2012 debates generated respectively 10.3 million[1], 7.2 million[2], and 6.5 million[3] tweets, all in a time span of approximately 90 minutes. Currently, only Twitter (the company) is able to make sense of such data by creating aggregations like the one shown in Figure 1. By establishing a correspondence between what was being said during the debate and the number of tweets per minute mentioning those words (and other debate-related hashtags), Twitter can quantify how moments during the live event affected the tweeting public. This technique was first applied by researchers during a debate for the 2008 US Elections [2]. A picture like the one in Figure 1 is a good way to summarize an event, especially its content. However, one might be interested not only in what is being tweeted, but also who is tweeting it and why.

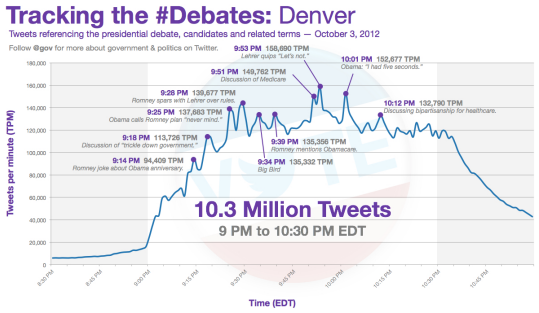

Figure 1: Image produced by Twitter to visualize the volume of tweets during the 1st Presidential Debate, Oct 3, 2012.

2. VISUALIZING THE BIG PICTURE

How can we get a more detailed dynamic of this real-time, event-specific conversation? Human curators, no matter how well trained, will hardly be able to keep track of the avalanche of data coming at a speed of 100,000 tweets/minute. One solution might be to use human computation: the independent decisions made by large groups of users instantaneously. In Twitter, this takes the form of retweets and favorites. In fact, if we were to look at the most retweeted

[1] http://blog.twitter.com/2012/10/dispatch-from-denver-debate.html
[2] http://blog.twitter.com/2012/10/twitter-at-town-hall-debate.html
[3] http://blog.twitter.com/2012/10/the-final-2012-presidential-debate.html

MSM '13, May 1, 2013, Paris, France

tweets of every debate, they capture unexpected moments that caused a stir, such as Mitt Romney stating he will cut funding for Big Bird. However, just because these tweets are the most retweeted, doesn't mean they necessary reveal the nuances of the conversation; they exhibit the usual drawback of large-scale recommender systems, the "popularity contest" effect. In an effort to put this human computation to a better use, we propose a new visualization technique that captures the dynamics of Twitter conversation during a certain event as evaluated by the co-retweeting behavior of the millions of users following the event. The data we collected during the debates shows that more than 50% of the tweets are retweets. However, the content being retweeted is much less, since many of the tweets will receive in the order of several thousand retweets (while a larger number will only receive one or two). We use this the relation between tweeters, retweeters, and the tweets being retweeted to create a co-retweeted matrix, that captures how the wide public views the participants in the discussion. This process is summarized in Figure 2.

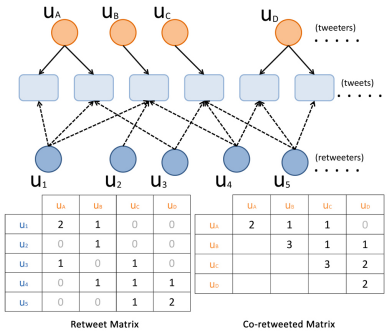

Figure 2: The process of generating the symmetric co-retweeted matrix. The main diagonal shows the number of users retweeting a certain account, and the other cells contain the number of times two accounts were co-retweeted.

3. DATA AND ANALYSIS

During the second and third presidential debates, we used the Twitter Streaming API to collect tweets containing specific keywords: debate, debates, debate2012, obama, romney, etc. In the moment that the conversation containing certain hashtags passes the threshold of 1% of the firehose volume, Twitter caps the number of tweets one can retrieve. For example, we were able to receive about 3,000 tweets/minute. However, we have been able to verify that our sample is representative of the most retweeted tweets.

Figure 3 shows the visualization of the co-retweeted matrix built with about 1.3 million tweets that we collected before, during and after the second debate, on Oct 16, 2012. The graph was created in Gephi, a network visualization tool[4]. We notice two very distinct groups, with only a few nodes bridging the gap between them. The colors of the graph make these two groups more apparent, and are based on a community finding algorithm, such that users within the a group are more closely connected to each other than those outside it. The size and darkness of each node is based on the Eigenvector Centrality algorithm, which ranks node based on their influence in the network.

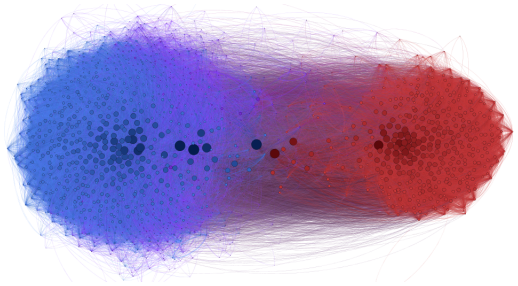

Figure 3: The co-retweeted matrix visualization for 1500 most retweeted accounts during the second debate. For an interactive version of this image, visit `http://cs.wellesley.edu/~sfinn/msm/oct16.html`

4. DISCUSSION

The visualizations for both debates[5] display two communities that break down strongly on political lines. The accounts in the blue group are largely liberal leaning politicians and popular bloggers (e.g. @barackobama, @thinkprogress), as well as popular news media accounts in the center of the graph between the two groups (e.g. @polifact, @huffpostpol). The red group consists of conservative accounts, (e.g. @glennbeck, @michellemalkin). Earlier research has demonstrated how political social media is polarized in the two political orientations [1], however, here the situation is different. The links are not created as a result of the actors actively connecting to each-other, but by how these actors are perceived by the public. The graph visualizes the political beliefs of the public at large.

During a real time event like the presidential debates, it may be difficult to find interesting content on Twitter. People use many hashtags, and there is huge amount of content being generated. Our visualization can help a user navigate this content by choosing to focus on users in the graph who she is familiar with, and then displaying similar Twitter users. Users who are tweeting actively during the debate might not tweet in such high volume outside the event, or tweet about content the user is not interested in. However, during the debate, she can find these users on the graph, and access the content they are generating.

5. REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] N. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *CHI*, pages 1195–1198, 2010.

[3] C. Pew Research. One-in-ten 'dual-screened' the presidential debate. `http://www.people-press.org/2012/10/11/one-in-ten-dual-screened-the-presidential-debate/`, 2012.

[4] `http:\\gephi.org`

[5] View Oct 22 debate at `http://cs.wellesley.edu/~sfinn/msm/oct22.html`