Peaks and Persistence: Modeling the Shape of Microblog Conversations

David A. Shamma
aymans@acm.org

Lyndon Kennedy
lyndonk@yahoo-inc.com

Elizabeth F. Churchill
churchill@acm.org

Internet Experiences, Yahoo! Research
4301 Great America Parkway, Santa Clara, CA 95054 USA

ABSTRACT

A microblogged stream is delivered over time, providing an ongoing commentary of topics, trends, and issues. In this article, we present two methods of finding temporal topics within these Twitter streams. Using a normalized term frequency, we demonstrate how an effective table of contents can be extracted by finding localized "peaky topics". Second, we find "persistent conversations" which have a lower general salience but sustain and persist over the tweet corpus, in effect the whispering conversation that lingers in the background. These methods are demonstrated on a Twitter corpus of 53,000 tweets and a second Twitter corpus of 1.1 million tweets; the methods are generalizable to apply to any normalized scoring metric across a temporal corpus. We propose our method's implications on social media research and systems from a textual and social network analysis perpective.

General Terms

Human Factors

ACM Classification Keywords

H.3.1 Information Storage and Retrieval: Content analysis and indexing—*Indexing Methods*; H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces—*Collaborative computing*

Author Keywords

microblogging, events, Twitter, inauguration, MTV, conversation, commentary, information, retrieval, IR

INTRODUCTION

Microblogging services have become a highly active forum for comments, opinions, and reactions—all of which convey social awareness and presence to a set of subscribers. Of the microblogging services, Twitter is of interest because of its scale and annual growth rate[1]. Owing to a short 140 character post size, it is easy to contribute messages from dedicated applications and web browsers on mobile devices, and desktops. Twitter usage has fueled the discussion of the "real-time web" while transforming commenting from the asynchronous synchronous to an almost concurrent conversation.

When tweets relate to an event, they can be used to understand the event's structure [7]. During sports events and live TV broadcasts, people cheer and react online while watching first hand. Within these Twitter streams, we wish to investigate the ongoing temporal conversation that exists as momentary topics of interest and longer trending conversations that are sustained and persist throughout the stream. In this work, we aim to find momentary topics, like a table of contents, as well as what is being discussed or perhaps whispered in the backchannel. The metrics we present can be generalized beyond the two Twitter test cases we employ; they can be easily utilized to examine any time delivered data/content stream.

BACKGROUND

To date, Twitter and microblogs have been largely studied from social perspectives. The text-based studies and visualizations, like Eddi [2], rely on information retrieval techniques [1, 5]. Vieweg et al. examined Twitter's usage during disasters and suggested some information extraction strategies [8]. With respect to social analytics in Twitter, friendship reciprocity and social diffusion has taken the general focus and fueled several studies [3, 4]; this work does not generally examine the text or identify the conversation. In these cases, the authors rely on active participation to a visible or known event. Our primary contribution describes how to identify an event stream's unfolding "table of contents" and the ongoing "background whispers" from the textual patterns of social activity from the active to the peripheral participants.

PEAKY AND PERSISTENT TOPIC IDENTIFICATION

The text of tweets can reveal a great deal about the structure and activity of the event contained in the stream. It also describes the relative level of interest that individual moments generate, like when everyone tweets "goal" during a world cup match. We believe that the temporal evolution of the textual content of tweets can point towards and semantically annotate important moments and eventually predict topics of on-going discussion and interest. To investigate this, we describe two metrics: *peaky topics* that show highly localized, momentary terms of interest and *persistent conversational*

[1] `http://blog.comscore.com/2009/04/twitter_traffic_explodesand_no.html` Accessed 8/2010.

CSCW 2011, March 19–23, 2011, Hangzhou, China.

topics that show less salient terms which sustain for a longer duration.

To mine text across these two metrics, we employ a simple term scoring approach similar to the well known *tf · idf* model [6]. In *tf·idf*, the salience of a term in a particular document is given as a function of the number of times the term appears within the document (term frequency, or *tf*) normalized by the total number of documents in which the term appears (inverse document frequency, or idf). Traditionally each tweet is a document and *tf · idf* would return a unique term score for each term *in each tweet*. This does not give an aggregate view of the overall term usage. We overcome this by creating an alternate *pseudo-document* composed of all the terms tweeted over a given time frame. In our work, the *tf* is the number of times the term occurs in this pseudo-document. The *df* is defined as we decribed above.

Peaky Topics

Peaky topics are terms which are particular to an exact window of time and not salient to other windows; they examine the frequency of various terms over a windowed time period. This begins by scoring terms according to their "window term frequency," $tf_{t,i}$, or the number of tweets containing term i within a given temporal window around time t. We normalize this value by the "corpus term frequency," cf_i, which we define as the total number of tweets containing term i across our entire collection. Using these two measures, we arrive at a "normalized term frequency" score as $ntf_{t,i} = \frac{tf_{t,i}}{cf_i}$ which can be intuitively described as the percentage of the total tweets containing term i that occur within the window around time t. For the purpose of our experiments, we set the size of the sliding window to be 5 minutes (± 2.5 minutes around t) and calculate normalized term frequency scores once for each slice across the entire corpus.

We expect that moments of interest will have terms associated with them that are highly frequent in the temporal vicinity of the event and relatively infrequent at other times. To automatically find such moments, we rank each term in the data set according to its *peakiness*, which we define as the maximum value of $\text{ntf}_{t,i}$ for term i. Intuitively, the *peakiest term* that we could possibly find would have a maximum normalized term frequency score of 1: all occurrences of the term fall within one window. On the other hand, non-peaky terms will have a uniform normalized term frequency score across all windows: the frequency of usage does not vary over time. If term i reaches a significant peak at time t, we can infer that there is a moment of interest at that time and that the term is a reflection of the content of that moment.

Persistent Conversations

In addition to the popular, momentary topic trends, we also wish to find *persistent conversations*: less salient terms and topics which are temporally sustained for a duration of time. We further expect that topics with sustained levels of interest will also be reflected in the temporal evolution of term usage on Twitter as well as find topics and issues which endure well beyond the related event stream. To automatically find such moments, we find the time $t_{\text{peak},i}$ at which the peak in the normalized term frequency score occurs for each term i. For terms with sustained interest, we suspect that the term will be seldom used before $t_{\text{peak},i}$ and then more frequently used afterwards. To evaluate this intuition computationally, we calculate the average values of $\text{ntf}_{t,i}$ for $t < t_{\text{peak},i}$ (pre-peak) and $t > t_{\text{peak},i}$ (post-peak) for each term. We score each term for its level of sustained interest by taking the ratio of the average post-peak score over the average pre-peak score. We then rank all terms according to their individual sustained interest scores.

EVALUATION

We will test our metrics with two datasets. The first data set is a representative sample of 53,712 tweets from the Inauguration of Barack Obama. This data was pulled from an API stream of the Twitter public timeline. The "data-mining stream" averaged 597 tweets per minute and was collected January 20, 2009 from 11:30 AM to 1:00 PM. This stream has since been deprecated by Twitter and is superseded by the "garden hose." The second data set is a more so complete sample of 1.1 million tweets from the MTV Video Music Awards (VMAs), which was acquired through a white-listed track feed. The VMAs dataset contains tweets from the 4 hour period between September 13th, 2009 at 8:30 PM to September 14th, 2009 at 12:30 AM. We will describe the two metrics in detail using the first data set. The second, larger MTV data set will be used to provide insights into how these metrics work at scale.

We pulled each sample with 30 minutes before and after each event to help identify the start and end of the program. Each data set had roughly the same percentage of mentions, @ symbols directly referring to another Twitter user, 23% during the inauguration and 28% during the VMAs. There were however differences in how many URLs were shared (15% vs. 3%) and how many retweets occurred (2% vs. 7%), inauguration to VMAs respectively.

Application: Obama's Inauguration

For the Obama inauguration, our dataset began at 11:30 AM and continued to 1 PM. The ceremony filled the 30 minutes from 12 PM to 12:30 PM. The 2-minute swearing in of the President occurred at 12:05 PM EST. Around minute 56 (12:25 PM), the inauguration speech concludes. In Figure 1, we show the normalized term frequency scores over time for the terms with the highest peakiness scores. Each of these terms distinctly reflect actual events in the inauguration proceedings. The terms "aretha," "yoyo," and "warren" reflect the appearances of Aretha Franklin, Yo-Yo Ma, and Rick Warren, respectively. The appearance of "booing" corresponds to the appearance of George W. Bush and a peak in "chopper" occurs when he departs via helicopter. "Remaking" is the highest-ranked of a cluster of terms that echo the content of Obama's address and "anthem" peaks as the national anthem is played.

Since we implemented our metric with unigrams, a single event topic can have multiple terms associated with it. For example "aretha," "franklin," "bow," and "sings" are four of the top-six overall peakiest terms, but each is reflecting the

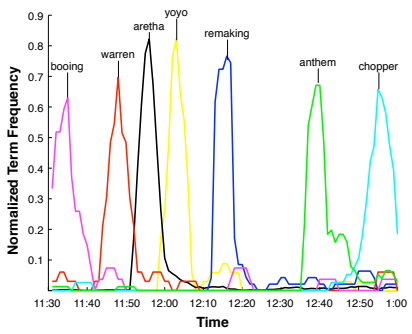

Figure 1. The top *peaky* term per window from the 2009 Inauguration of Barack Obama.

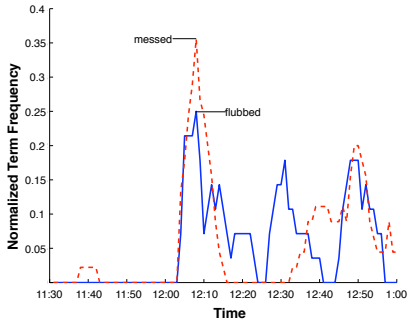

Figure 2. Top two terms of *persisted* usage from the Obama Inauguration of 2009. These terms are relatively infrequent before the occurrence of an event of interest. They peak in frequency around the event and continue to be used for a period of time afterwards.

same event: Aretha Franklin's performance and the bow on her hat. We had to correct this by removing such duplicate event labels—skipping terms that are highly correlated ($p < 0.05$) with a higher-ranked term.

In Figure 2, we show the two terms that we find to have the highest level of persisted interest: "flubbed" and "messed." Both are related to Chief Justice Roberts mistakenly switching the order of a few words while administering the oath of office to President Obama. Both terms are virtually never used before the oath incident and then suddenly peak around the event. However, unlike the peaky terms shown in Figure 1, they continue to be used for a great deal of time after the event. This particular conversational topic received a great deal of media attention in the days following the inauguration, which may have been predicted almost instantaneously by this tweeting behavior.

Application: Video Music Awards

In the 2009 MTV Video Music Awards, we begin by examining the end of the "pre-show" and the awards show itself, which began at 9 PM Eastern time and concluded at 11 PM. In this show, there was a particular unexpected incident. While Taylor Swift was receiving her award for Best Female Music Video, another performer, Kanye West, jumped up on stage, took the microphone from her to say he felt another video was better. This created much controversy and the

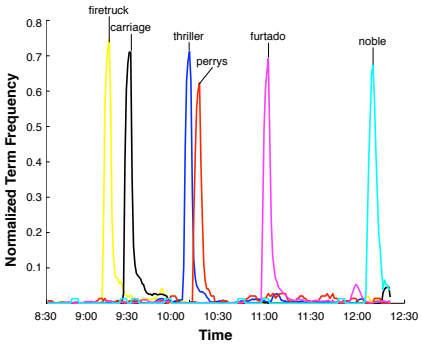

Figure 3. The top *peaky* term per window from the 2009 MTV Video Music Awards.

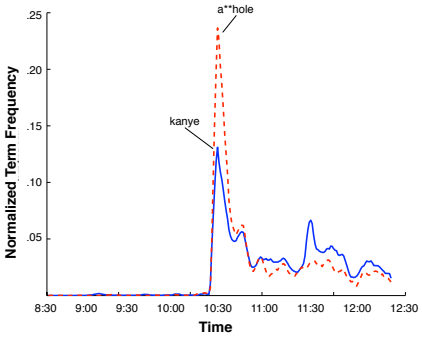

Figure 4. Top two terms of *persisted* usage from the 2009 MTV Video Music Awards.

effect was evident on Twitter[2]. This event occurred approximately 26 minutes into the program.

We applied our analysis of normalized peaky term frequency over the VMA dataset to discover peaky terms and points of interest. Some examples of the discovered terms are shown in Figure 3. Again, we see a tendency for activity on Twitter to reflect the events that are unfolding on the screen. Towards the beginning of the event, we see appearances of terms like "firetruck" and "carriage," which are in response to the vehicles in which certain artists are arriving to the event. During the primary awards show broadcast, we see a tendency to reflect which presenters or performers are on stage and perhaps which song they're performing: "thriller" peaks as Janet Jackson performs a tribute to Michael Jackson, "perrys" surfaces while Katy Perry and Joe Perry perform together, and "furtado" appears when Nelly Furtado presents an award. Towards the end of the event, we see a peak in the term "noble" which occurs as Taylor Swift is allowed a second chance at her acceptance speech after being interrupted by Kanye West. "Classiest" and "gesture" are other peaky terms that are highly associated with this particular point in time.

West's interruption of Swift is reflected in our analysis of the terms with the most persisted interest, shown in Figure 4. As West grabs the microphone shortly after 10:30

[2] `http://content.stamen.com/kanye_west_is_an_a__hole_and_other_twitter_moments` Accessed 8/2010

PM, the usage of the term "kanye" begins to surge on Twitter, along with other characterizations of West, particularly calling him an "a**hole." This persisted interest was later echoed in ongoing discussions of the moment and various incarnations of Internet memes that appeared in the following week. Again, the initial emergence of sustained term usage trends on Twitter predicts discussions of moments of interest that will persist well beyond the immediate occurrence of the given event.

Discussion

From these two metrics, we were able to identify topic terms which are momentarily salient (peaky) and topic terms which are conversationally persisted. Both of these metrics could benefit with n-gram analysis, but would still need the correlated term phrases removed to identify uniqueness amongst the topics. Both metrics preformed well at scale. In particular, the trending conversations metric identified salient topics by using a $tf \cdot idf$ approach specifically modified for measuring group conversations.

Our method counts the number of tweets that a term appears in, rather than raw total frequency of the term, since we focus on gauging the number of *people using a term* at a given time. This avoids biases that might be caused by a single user repeating a term many times in one tweet, which we did observe. The method also uses corpus frequency as a substitute for document frequency, which yields a normalized score that is more suitable for further analysis and comparison. As a consequence, our method, like $tf \cdot idf$, is also sensitive to very infrequent terms. This can be avoided by removing terms below a certain frequency threshold.

In the trending conversations, the usage of @mentions, references directed towards other users, in tweets containing these two terms also evolves over time. If we separate the tweets containing "flubbed" or "messed" into two groups: those around the time of the oath (before 12:15) and those after the oath (after 12:15), we see a distinct difference in the type and level of conversation.

The initial set of tweets around the time of the oath simply note and react to the mistake. For example:

> (12:05) **Bastille71:** OMG - Obama just messed up the oath - AWESOME! he's human!
>
> (12:07) **ryantherobot:** LOL Obama messed up his inaugural oath twice! regardless, Obama is the president today! whoooo!

Meanwhile those that follow in the ensuing hour afterwards are further conversations about the incident and contain instances of people discussing the oath and correcting (sense-making) each other:

> (12:46) **mattycus:** RT @deelah: it wasn't Obama that messed the oath, it was Chief Justice Roberts: `http://is.gd/gAVo`
>
> (12:53) **dawngoldberg:** @therichbrooks He flubbed the oath because Chief Justice screwed up the order of the words.

Only 7% of the tweets in the first set contain @mentions, compared to 47% in the second set. During the VMAs we observed a similar yet less agressive pattern as there was less sense-making needed with regards to Kanye's surprise appearance. Removing references to @taylorswift13 and @kanyewest, 12% of the first set contained mentions compared to 19% in the following tweet set. We expect retweets (tweets explicitly repeated by other users) could be handled by a calculated $\pm$ scalar on that term's score, depending on application.

FUTURE WORK

The textual content of tweets can reveal a great deal about the structure and content of the event as well as the relative level of interest that individual moments generate. We have begun to indentify patterns in common between events that maintain interest over time (sustaining or periodic) versus events that are moments that do not persist (or repeak) over time. In particular, we believe that the temporal evolution of the textual content of tweets can point towards and semantically annotate important moments and predict topics of on-going discussion and interest.

We highlight that our computations are calculated for each minute of the event. We then used the temporal evolution of these scores for each term to classify terms as 'peaky' or 'persistent.' We believe two metrics can be applied towards any normalized scoring method and are easily applicable to existing studies of microblog usage [8, 2]. We believe these metrics can be can be applied to social network analysis, in particular network centrality, to identify people in the communication graph as temporally salient actors.

REFERENCES

1. M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 491–498, New York, NY, USA, 2008. ACM.
2. M. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. Chi. Eddi: Interactive Topic-based Browsing of Social Status Streams. *ICWSM*, 2010.
3. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
4. B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.
5. M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA, 2006. ACM.
6. G. Salton and M. McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc. New York, NY, USA, 1986.
7. D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: Understanding community annotation of uncollected sources. In *WSM '09: Proceedings of the international workshop on Workshop on Social Media*, Beijing, China, 2009. ACM.
8. S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088, New York, NY, USA, 2010. ACM.