A Qualitative Examination of Topical Tweet and Retweet Practices

Meenakshi Nagarajan, Hemant Purohit, Amit Sheth
Kno.e.sis, Dept. of Computer Science and Engineering
Wright State University, Dayton, OH
{meena, hemant, amit} @knoesis.org

Abstract

This work contributes to the study of retweet behavior on Twitter surrounding real world events. We analyze over a million tweets pertaining to three events, present general tweet properties in such topical datasets and qualitatively analyze the properties of the retweet behavior surrounding the most tweeted/viral content pieces. Findings include a clear relationship between sparse/dense retweet patterns and the content and type of a tweet itself; suggesting the need to study content properties in link based diffusion models.

Introduction

Twitter's popularity in harnessing real-time traffic, enabling large-scale information diffusion and creating tangible effects on economies and societies is well known today. Minutes after President Obama's address to Congress on healthcare, Twitter showed an avalanche of tweets about the outburst from Joe Wilson. Twitter's influence was also apparent following the terrorists attack in Mumbai and in the civil reaction to the Iranian elections.

A user post on Twitter (i.e., a tweet) comprises of the poster's unique identifier, a time stamp reflecting when the tweet was posted and a 140 byte long content itself. Users of Twitter have directed 'follower' connections with other users of the site that allows them to keep track or 'follow' those users. Members can post tweets, respond to a tweet or forward a tweet. Replies to any tweet are directed to a user (not the conversation thread) utilizing the @user reference while retweets are means of participating in a diffuse conversation. The content of a tweet typically also contains text and hashtags (e.g., #iranelection) that indicate explicit topic categorization and links to other multi-media content that promote spread of information from all over the Web. Tweets are generally available as feeds from follower networks and also via a searchable interface.

One of the highlights of Twitter is its *retweet* functionality that allows members to *relay* or *forward* a tweet through their network. In a recent empirical analysis of tweets [boyd, d.], the authors presented various conventions and styles of retweeting prevalent today. They also noted the emergence of retweeting as a conversational practice in which 'conversations are composed of a public interplay of voices that give rise to an emotional sense of shared conversational context'. This is especially true of real-world events where a community volunteers participation and engages in topical conversations. While it has been argued that, as with link-based blogging [Marlow, C.], retweeting holds immense potential for viral marketing and content sharing, the wide-spread prevalence or mechanics of this practice for topic-specific discussions has not been documented. Understanding properties of information diffusion around an event gives us cues into the dynamics of the community that rallies around a particular cause. This study is a first step in that direction. We focus our study on topical tweets generated by communities that gather on Twitter around real-world events. We analyzed a total of 1677978 tweets pertaining to three different events - the Iran Election (IE), the Health Care Reform debate (HCR) and the International Semantic Web Conference (ISWC). Each of these events is of varied social significance, attracts different populations, spans different time periods and lengths of time and therefore represents a wide variety of twitter activity. In the next two sections, we characterize the data used in this study and present our findings on the observed tweeting and retweeting behavior.

Data Collection Methodology

The data used for this work was collected as part of a social Web application, Twitris [Nagarajan, M], that presents spatio, temporal and thematic summaries of user tweets surrounding an event. Tweets were crawled with Twitter's search API using an initial seed of manually compiled keywords and hashtags relevant to the event. For a keyword k, we crawl all tweets that mention k, K, #k and #K. The seed list of keywords and hashtags is kept up-to-date by first automatically collecting other hashtags and keywords that frequently appear in the crawled tweets and then manually selecting highly unambiguous hashtags and keywords from this list. For example, in the ISWC conference, we started with two keywords and their hashtags – 'iswc' and 'iswc2009' and ended up with a final seed list of around 50 keywords that reflect discussions surrounding the event – for example, 'sdow2009', 'linkeddata', 'semanticweb' etc. We avoid the query drift problem by placing a human in the loop to ensure that ambiguous keywords like 'nyt' are not crawled outside of context but only in combination with a contextually relevant keyword, for example, 'nyt' and 'linkeddata'.

Data crawl was performed at fixed time intervals depending on the nature of the event. A fairly focused event like the ISWC conference that has a small participant base was crawled every minute while the Iran Election event that attracted world-wide attention was crawled every 30 seconds. For every issued query, the Twitter search API responds with 1500 tweets. Crawling at regular and frequent intervals allows us to make an assumption that the data collected is a close approximation of the actual population of the tweets generated for the event in that time period. We also collected poster location (from the user's profile) and timestamp information associated with the tweet. Data statistics are shown in Table 1.

Table 1 Data Statistics

Event	#Tweets	Date Range	Data Source (top3)	#Unique Posters
HCR	1163687	Aug13-Dec22'09	USA, Canada, Mexico	223274
IE	508959	Jun4-Jun30'09	Iran, USA, Canada	142831
ISWC	5332	Oct 19-Nov 8'09	USA, Canada, UK	2437

Macro-level Summaries of Tweets and Posters

In our first study we tried to get a sense of what types of tweets dominate community activity surrounding the three events. We found fairly strong allegiances to topics via the indication of hashtags in the content. Approximately 48% of tweets in the HCR dataset, 68% in the IranElection and 52% of tweets in the ISWC dataset had one or more hashtags present in them. This shows that a majority of users do care about their voice being heard, allowing it to be categorized and found. Since we use both keywords and hashtags in obtaining the data, this allegiance is not biased by the community-classified tweets alone.

To get an indication of user engagement in these very specialized communities, we calculated the following:

1. proportion of tweets that made references to other Twitter users utilizing the @user handle (implying a directed conversation, i.e., a reply or reference),
2. proportion of tweets that were retweets (those containing one of the several explicit RT conventions prevalent today, e.g., RT @user, rt @user, retweet @user etc,. also listed in [boyd, d.]), and
3. proportion of tweets posted by users without indication of retweeting or making reference to others.

Table 2. Proportion of Tweet types

Tweet type	HCR	IE	ISWC
1. % directed conversations	12%	8%	23%
2. % retweets	27%	44%	24%
3. others	61%	48%	53%

We found a user engagement pattern that was consistent across all three events. Overall, users engaged in fewer directed conversations and more retweet engagements, for example, 8% vs. 44% in the Iran Election dataset (See table 2). This is not very surprising since formulating a reply or making a reference involves more cognitive overhead compared to forwarding of a tweet. However, the ISWC dataset suggests that people equally engage in both types of tweeting behavior. While this is a smaller dataset, it is likely that the familiarity between members of that community and the narrow focus of the event dictated such a level of engagement. In all three cases, the proportion of singular tweets that were not retweets or replies/references is the largest. This is to be expected because not every tweet posted on Twitter catches the attention of a community to warrant a response or a retweet. These observations have to be interpreted in light of the fact that we would have missed topical tweets that do not use the keywords or hashtags that we used for the crawl.

To get a sense of the active population within these datasets, we studied the user base across three different types to reflect their engagements:

- Active posters: users who posted the most number of tweets for that event.
- Popular mentions: users that received most number of references using the @user handle but were not a part of RTs. This metric is a soft indicator of popularity since the community engages in direct conversations with these users.
- Popular retweeted authors: users who were most retweeted using one of the many conventions. This is an indication of how authoritative or pertinent the community finds these users' posts.

We found a common, rather intuitive pattern across the top 10 users in the three datasets. News and marketer profiles were the most active posters in terms of the number of tweets they generated. Users 'pr health', 'PRNhealth', 'SemanticBot' and 'SemanticNews' appeared in the top 5 tweeters in the HCR and ISWC events. When it came to retweeting or mentioning users, the community favored individual posters (those that were not news or marketer profiles). The names of the individual posters are not presented owing to a privacy concern raised by one of the reviewers of this paper.

The Anatomy of Popular Tweets

Table 2 gives us a sense for how much of the data can be attributed to the *retweeting* behavior, i.e., what percentage of tweets were reposted or forwarded in the community. To understand finer nuances of retweeting, we decided to focus on the most popular or viral tweets.

Methodology: We extracted the top 10 most tweeted / viral tweets in every event dataset and looked at their retweet patterns. Tweets that are copies of each of these 30 tweets were gathered based on a high content similarity (Levenshtein string similarity of 0.75) after ignoring user references, hashtags and hyperlinks. This also allowed us to capture variations of the same tweet. For example, the following were grouped as indicating the same content.

Twitition: Google Earth to update satellite images of Tehran #Iranelection http://twitition.com/csfeo @patrickaltoft
Just signed petition 'Google Earth to update satellite images of Tehran ' http://301.to/23o

Top 3 tweets: Here we show the top 3 most viral tweets (given space restrictions) in their respective datasets. One interpretation for why these tweets are viral is Twitter's popularity in drawing real-time traffic and facilitating conversations around trending topics [Stross, R.].

Health Care Reform Debate
1. Powerful video from @MoveOn and R.E.M. about the real lives at stake in the health care debate. http://bit.ly/UVqZl #publicoption
2. Join @MarkUdall @RitterForCO and @BennetForCO to support an up or down vote on the public option http://tr.im/Cm2u
3. Tell John Boehner that you are one of millions of Americans who supports a public option http://dccc.org/tellboehner #p2 #publicoption #hc09

Iran Election
1. #iranelection show support for democracy in Iran add green overlay to your Twitter avatar with 1 click http://helpiranelection.com/
2. Twitition: Google Earth to update satellite images of Tehran #Iranelection http://twitition.com/csfeo @patrickaltoft
3. Set your location to Tehran and your time zone to GMT +3.30. Security forces are hunting for bloggers using location/timezone searches

ISWC
1. NYTimes linked data now out at http://bit.ly/lghDo potentially very useful and all mapped to DBPedia, Freebase etc #iswc2009
2. Talis will be hosting a series of Open Days in Birmingham: http://bit.ly/4iukyo. Get a hold of data and learn about SPARQL.
3. The Semantic Web is the future of the Internet. Always has been. Always will be.

At the outset, we noticed that only 20% of the retweets of the top 10 tweets in the Iran Election and Healthcare dataset followed the explicit retweet syntax and included an author attribution. On the other hand, 78% of retweets of the top 10 tweets in the ISWC dataset contained author attribution information. We did not find any correlation between the original posters of these 30 tweets and the presence of or lack of attribution. We turned to the content properties to shed some light on this behavior. Here, we present results from a qualitative examination of the tweets' content properties and its correlation with the phenomenon of author attribution.

Content Properties: Although not an exhaustive list, all 30 popular tweets appeared to encourage one of the following; observations that are also in line with those made by [boyd, d.]:

- call for some sort of social action: *"show support for democracy in Iran add green overlay to your Twitter avatar with 1 click"*.
- collective group identity-making: *"Join @MarkUdall @RitterForCO and @BennetForCO to support an up or down vote on the public option"*.
- crowdsourcing: *"Tell John Boehner that you are one of millions of Americans who supports a public option"*.
- information sharing: *"Powerful video from @MoveOn and R.E.M. about the real lives at stake in the health care debate."*

Two of the three authors classified each of the 30 tweets into one of the above categories with complete agreement in the classification. 60%, 40% and 90% of the top 10 tweets in the Iran Election, Health Care and ISWC dataset classified as information sharing tweets, while the rest classified as one of the other types.

For every tweet and its copies found in our dataset, we plotted the retweet connections between the authors of these tweets as a directed network/graph. Every node is a unique author who posted the tweet and the directed edges are *retweet* references made by one author to another. If user A retweets user B, an edge is drawn from node B to node A; indicating the direction of information diffusion. Observations and illustrations provided here are for retweet patterns over the entire time period that the tweet was visible on Twitter. Recall that nodes are authors that posted the tweet while edges imply one author retweeting another.

Finding 1 - Sparse Retweet Networks: All the popular tweets that categorized under the first three types of tweets mentioned above (call for action, crowdsourcing or collective group identity-making), generated sparse retweet graphs. In other words, although it was obvious that the content was being retweeted, re-posted or copied, author attribution for these types of tweet was absent.

Figure 1-A1. shows an example of the retweet network for a 'call for action' tweet - *"Join @MarkUdall @RitterForCO and @BennetForCO to support an up or down vote on the public option http://tr.im/Cm2u"*. Of the total 498 occurrences (copies and variants) of this tweet, only 34 explicit retweet/attribution edges were present, and the largest connected component consisted of only 10 nodes. The corresponding follower graph (Figure 1-A2.) for this tweet (how the authors were connected by follower links) was however well-connected; implying that people did (possibly) see these tweets from their network but did not feel compelled to credit the sender or the original author. Note that we cannot tell if the users followed each other before or after they posted the tweet.

Among all the popular tweets of the type 'call for action', 'crowd-sourcing' or 'collective group identity-making' across the 3 events, only 5% of the tweets contained any attribution information in them. We also studied the networks for such tweets on a day-to-day basis but did not find any patterns of attribution decay that could be linked to the posters or the content/tweet variations.

Potential reasons for sparse attribution: In our analysis of retweet patterns for this class of tweets, we found several reasons why author attribution might be sparse.

1. Typically, tweets that make a 'call for action' do not credit a person. Consequently, users do not feel compelled to pass on credit to a person who acted as a messenger.
2. Familiarity among users seemed to play an important role in sustaining author attribution. This was especially prominent in the case of retweet behavior patterns in the ISWC dataset where it is very likely that posters have an offline relationship in addition to their Twitter connection. Among the top 10 tweets from the ISWC dataset, 78% of the tweets contained author attribution.

3. Viral tweets of this type rely on a community gathering around a cause. Consequently, people's window to such tweets come from various sources – multiple people in the follower network, trending topics, elsewhere on the Web etc. It is possible that the user did not see the tweet from his network at all and hence does not attribute the sender.
4. As also noted by recent work on studying motivations behind retweets [boyd, d.], a potential reason for attribution information to not sustain could be that some users are trying to make space for their content and therefore losing attribution details.

Finding 2 - Dense Retweet Networks: In contrast to the previous class of tweets, we found that tweets sharing information (e.g., contained hyperlinks to informative posts, videos, images) generated a denser retweet/attribution network. Figure 1-B1. shows an example of an information sharing tweet *"Iran Election Crisis: 10 Incredible YouTube Videos http://bit.ly/vPDLo"*. For a total of 1399 tweets of this content, there were 949 retweet edges between the nodes and the largest connected component consisted of 778 nodes. Among all the popular tweets of this type ('information sharing'), across the 3 events, 79% of the tweets contained author attribution information.

Discussion and Conclusion

We observed the aforementioned sparse and dense retweet patterns to be true across the top 30 tweets from the 3 events. The patterns are consistent across the events despite the fact that the events were varied in the population they attracted and the goals of the communities. This categorization is certainly not exhaustive but suggests an important finding - the content being tweeted plays a key role in what an explicit retweet network will look like and in many cases, whether it will be traceable at all. General properties of tweeting practices were also consistent across the events and shed some light on topical tweeting behavior.

Since this study, there have been some changes made to the retweet functionality support via Twitter and other third-party interfaces to the medium. While this is something to keep in mind while interpreting the results, we believe that the suggestive relationship between the tweet type and its retweet pattern will contribute to the study of link-based diffusion models. In our ongoing and future work, we intend to use these qualitative results in quantitatively verifying the role of content properties in the virility of a larger sample of viral tweets. For example, in our recent experiments over 300+ viral tweets, we are seeing significant correlations between impersonal pronouns and verbs (typical of making a call for action) in tweets and the sparse attribution networks they generate.

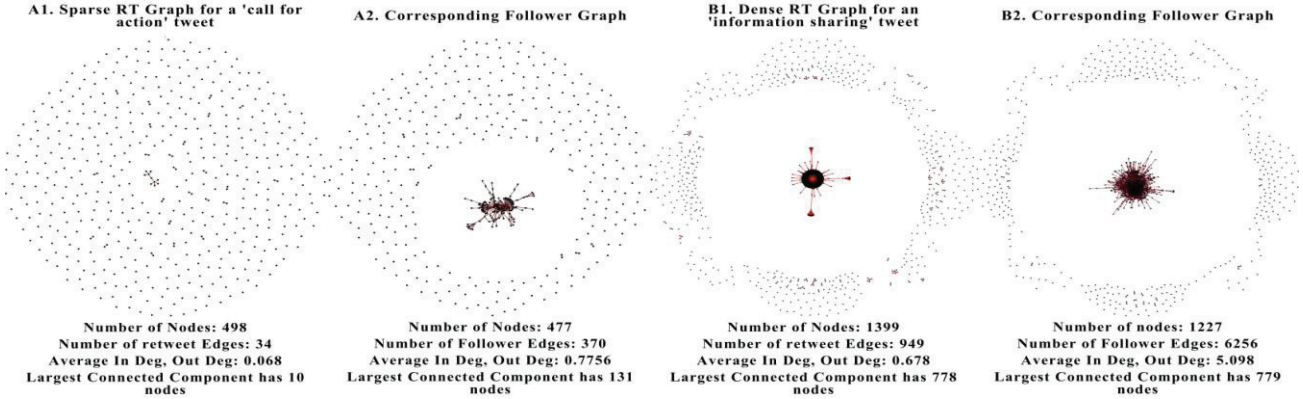

Figure 1. Graphs showing sparse (A) and dense (B) RT networks and their corresponding follower graphs for 'call for action' (A) and 'information sharing'(B) type of tweets respectively.

Acknowledgements

We would like to thank the reviewers of this paper for their comments and members of Twitris, Knoesis Center for providing the infrastructure for data collection. This research was supported in part by a Microsoft External Research award under "Beyond Search – Semantic Computing and Internet Economics program" and by NSF Award#IIS-0842129, titled "III-SGER: Spatio-Temporal-Thematic Queries of Semantic Web Data"

References

boyd, d.; Golder, S.; and Lotan, G. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *HICSS 43. IEEE: Kauai, HI.*

Marlow, C. 2005. The Structural Determinants of Media Contagion. Ph.D. Diss., MIT Media Lab.

Nagarajan, M.; Gomadam, K.; Sheth, A.; Ranabahu, A.; Mutharaju, R. and Jadhav, A. 2009. Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences. In *WISE'09: In Proceedings of the tenth international conference on Web Information Systems Engineering,* 539-553.

Stross, R. 2009. Hey Just a Minute (or why Google isn't Twitter). *New York Times*.