ReFluence: A Real-Time and Historic Visualization Application for Twitter Conversations

Ramine Tinati, Leslie Carr

University of Southampton, United Kingdom
rt506@ecs.soton.ac.uk

Abstract

Social networks provide a new and exciting way for individuals, businesses, organizations and governments to create and share information. Specific social networks such as the popular micro-blogging social network site, Twitter, provide individuals with an opportunity to disseminate information to a potentially global audience. In this paper we describe the ongoing development of ReFluence, which has been developed to visualize Twitter streams, providing a historic and real-time visualization of the growth of Twitter conversations between users, based upon the networks that form through the retweet feature. In addition to this, ReFluence also provides a way to identify and classify different users based on their Twitter behavior.

Introduction

In recent times, social media has grown to be an important medium to inform and disseminate news and information across a global audience. Microblogging, popularized and dominated by Twitter offers Web users the ability to produce short, 140 character messages, potentially visible to over 100 million users; with a recent record message throughput measured on Twitter now totaling 140 million tweets a day (Twitter, 2011). By tapping into this large amount of data, a number of tools and applications have been created to provide societal benefit, including identifying public health issues (Paul & Dredze, 2011), event detection (Weng & Lee, 2011), and even political uprising (Zhou & Kong, 2010). However, the enormous amount of data presents a problem, how can the valuable information be filtered from the background noise, which reportedly accounts for up to 40% of the total message traffic (Pear Analytics, 2009).

To some extent the Twitter architecture does allow for messages to be 'filtered' by topic, using the feature of hashtags (#), but this basic filtering soon becomes defunct when the number of tweets using the hashtag becomes too large; worsened when a hashtag becomes a 'trending topic', simply meaning that the hashtag has been mentioned in a large proportion of the total Twitter message stream, which then gets listed on Twitter's homepage.

In order to examine the possibilities of visualizing the network of communications that occur within a specific topic, we set out to develop an application that enables the conversations between users to be visualized both in real-time, and via a collection of saved tweets. In this paper we present ReFluence, a Twitter conversation visualization application which provides a way for the network of communications between users to be examined and also, based on a number of filters, provides a way to identify and classify different users within the network.

This paper will first describe the Twitter service, providing an overview of the possible conversation network structures that can be modeled, and then will describe the architecture and functionality of ReFluence.

Twitter

The Twitter service offers users the ability to post 140 character messages – known as tweets – to their own timeline of activity, which can be public or private. As part of the user model (and user experience), users can 'follow' each other, which creates a directed link between two users, as shown in Figure 1. The 'following' of a specific user results in the followed user's tweets appearing in the 'following' user's personal timeline. Depending on the followed user's notification settings, they may be told that they are being followed by a specific user, but it does not require them to 'follow' them back. This therefore results in a directed network of users (nodes) and followings (edges).

Twitter also enables users to write tweets with a 'screenname' of a Twitter user included as part of the message – performed using the @ symbol. As a result of this, the user that has been ‘mentioned’ in the tweet will view the tweet in their timeline. This therefore produces another network structure of users (nodes) connected together by tweets (directed edges), as shown in Figure 1. This sharing of the tweet is also altered by the network of user’s followers, for example, if user A follows users B and C, and B makes a tweet mentioning C, consequently, that tweet will also appear on user A timeline.

Similar to the ‘mention’ function, the ‘retweet’ function allows users to copy tweets from the timeline of the users they follow and add them to their own timeline. Twitter defines the retweet function as:

“A Tweet by another user, forwarded to you by someone you follow. Often used to spread news or share valuable findings on Twitter”.

As a result of this, followers of the user that has retweeted will see the retweet (and a link to the retweeted user) on their own timeline. This therefore creates a network where users (nodes) are connected by retweets (directed edges), but through a reverse directed graph, as shown in Figure 2. For example, if U2 retweets U1, then U3’s timeline will become populated with the tweet U2 has retweeted.

Finally, Twitter also offers users the ability to send ‘Direct Messages’ between users, as long as the user that sends the message is being followed by the message recipient. These messages do not appear on any public timeline and are only visible by the users involved.

In addition to the different ways Twitter users can create connections with different users based upon the syntax used within the tweet, Twitter also provides a way for Tweets to be ‘tagged’ via the use of the hashtag (#) features. For instance, if a number of tweets contain the string #foo then performing a search on via the Twitter API or the Twitter website for #foo (or foo) will return any all tweets containing that string. This provides a useful way for tweets to be categorized and searchable, and enables users to take part in conversations about specific topics. If a hashtag becomes used in a large number of tweets, these often become trending topics, and are made visible on Twitter’s homepage.

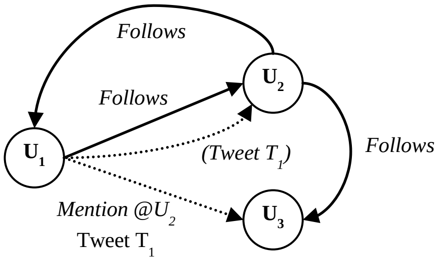

Figure 1 ‘Following’ and ‘Mention’ Network Structure

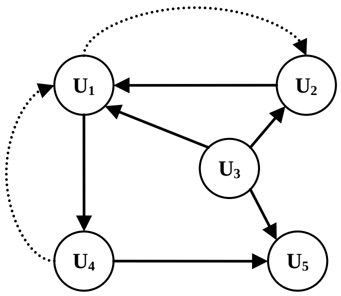

Figure 2 Twitter Retweet Functionality

ReFluence

The development of ReFluence was based upon the desire to be able to visualize a growing network of communications within the Twitter service. It was also part of a joint research venture with Edelman, a professional public relations company with an interest in understanding the capability of social media. The original aim of ReFluence was to provide an application which could provide a ‘playback’ the conversation between users, and also to provide a way for real-time discussions to be displayed.

The initial requirements of ReFluence was to provide a visual display of the growth of a specific Twitter conversation, which could be the based on either the tracking of retweets between users, or mentions between users. In addition to this, the temporal growth of the conversations could be paused, examined at specific points in time, thus providing a way to examine how the network of communication grows.

To restrict the visualization to users discussing a specific topic of interest, the Twitter hashtag feature provides a method limit the tweets within the conversation. The visualization also would enable the communications between users to be traceable, including details regarding the users within the conversation, and the tweets that were made.

An additional requirement, influenced by studies exploring possible methods to filtering and identifying important users in a Twitter conversation (Chu, Gianvecchio, & Wang, 2010) (Welch, Schonfeld, & He, 2011), the requirement was also added to enable identification of users based upon a range of criteria, such as their number of retweets within a conversation. This was explored further with Edelman, who had previously developed the *Topology of Influence* (TOI) (Bentwood, 2008) (Hargreaves & Davies, 2011), a theoretical classification model based upon their professional experiences within the PR industry.

Classification of User Roles

Based upon Edelman’s TOI, the requirement was added to be able to identify and distinguish users based on four different criteria: An *idea starter* – someone who has their tweets retweeted by a large number of people. An idea starter is calculated by finding the sum of all the retweets of a user divided by the minimum retweet number, if this is greater than one, then they are classified as an idea starter. An *amplifier* – a user who is the initial person to retweet a tweet, only if the tweet is part of a retweet chain. A *curator* – a user who retweets two or more idea starters. Finally *commentators* – a user who does not meet the above requirements, but have actively retweeted within the harvested data.

Modeling the Retweet Function

The retweet function is a user operated feature in which a user has to actively press the 'retweet' button, which then automatically creates an underlying directed edge from the user who retweeted to the original user. This provides a method to track the path of user messages, thus helping construct the network graph required.

The approach taken was to model the number of retweets that a specific user needed to be classified as influential; similar to the work of Cha (2010) and Welch (2011), we argue that the greater the retweets that a user incurs is a suitable indication of their influence – the retweet metric.

In order to determine if a user is influential, we set a value, defined as RT^{min}, which is the minimum number of retweets a user is required to have in order to be classified as influential. This therefore provides a method of examining the number of influential users depending on the number of retweets that they have achieved. Based on this and in combination with the previous research discussed, a model has been created which aims to understand the networks structures that resulted from the retweet functionality.

The retweet functionality provides a way to examine the flow of information between users over a large network. The hashtag (#) functionality provides a way of 'labeling' or 'tagging' a tweet to a specific topic or event. This enables specific searching for tweets relating that hashtag to be identified; useful for users and Twitter. The hashtag is an important feature to help identify influential users within specific 'topics'.

Thus, combining the method of calculating whether a user is influential using the RT^{min} value, with a network graph restricted to messages relating to a specific hashtag provides a way of examining the effects that retweeting have on Twitter's network structure.

Implementation of the Model

For the Twitter network graph to be constructed, the harvested data requires transforming into a network graph of nodes – representing users, and directed edges – representing retweets between users, as shown in Figure 3.The main data model that underpins the implementation builds a node and edge representation of all the tweets provided by the harvested data. Furthermore, as a result of the methods used to collect the data, it is also possible to produce a graph of retweets messages against a timeline of activity. As Figure 3 illustrates, there the 'Retweet Edge' will consist of the source (the user who made the original tweet), the target (the user who has retweeted the tweet) and also the tweet that has been retweeted.

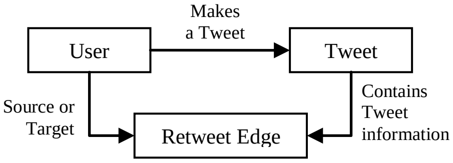

Figure 3 Data Model Overview

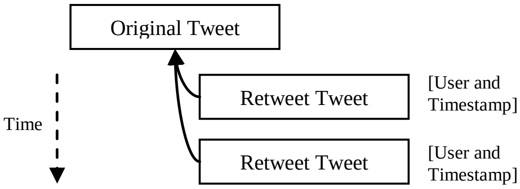

Figure 4 Retweet Chain

An edge, directed from the user that has retweeted to the user that original made the tweet, makes it possible for the model to calculate the total number of retweets that a user has received within the given dataset. This therefore provides a way to determine whether the users are classified as an influential user.

Including the tweet information for each of the edges is important for tracking the conversation flow within the harvested data. As the model is interested in finding the flow of conversation (which in this case, is the retweet flow), the model requires some way of building up a record of retweets. Figure 4 is a representation of the way ReFluence models the cascade of a retweet, over the entire conversation timeline. Calculating the length of a retweet chain and who was first to retweet provides another way to identify different user within the Twitter conversation.

Data Sources

To model Twitter conversations network graph, a dataset is required that contains all user and tweet information, bounded by a specific hashtag. The Twitter API provides a service to search for tweets based on a specific hashtag, which returns a list of tweets, with their corresponding metadata including: tweet identifier, user identifier and timestamp of tweet.

To cater for the history playback of a Twitter conversation, the approach taken was to use a custom designed EPrints plug-in[2], which enables a large volume of hashtag specific Twitter data to be collected. The tweet collection can be then exported in the appropriate format, providing a dataset which contains a chronological timeline of tweets corresponding to a specific hashtag.

[2] http://bazaar.eprints.org/161

For the real-time visualization of hashtag specific tweets, ReFluence requires access to the Twitter API to search for Tweets based upon a given search term or hashtag. This presents a number of problems, mainly due to the throttling of connections to the Twitter API, limited to a maximum 350 per hour (150 if not using the OAuth functionality[3]). To achieve the best possible representation of the real-time tweets stream, the solution was chosen to call the Twitter API every 30 seconds and use paging techniques to avoid the loss of a potential tweets. The Twitter search API provides a maximum of 100 returned results per call, which therefore provided a maximum of 12000 tweets to be collected per hour; the collected data could then be saved again for future playback.

Implementation

Based upon the requirements and architecture described, an application was built consisting of three components: (1) the model – as described in the previous section, (2) a front end GUI – which visualizes the growth of the Twitter retweet graph based on the timestamps within the dataset, and (3) a statistics simulator – which provides a visual output of the influence model applied to the loaded dataset.

The graph visualization front-end utilizes the JUNG[4] toolkit and draws a retweet graph using JUNG's Fruchterman-Rheingold layout capabilities. The main interface presents an updating graph of the retweet messages found within the harvested Twitter data. A timeline of retweet messages between users is drawn, presenting the user with a demonstration of the retweet conversation growth, at any point during this playback of events; it can be paused, allowing the user to find more additional information about a chosen node. The user is also presented with a slider, which sets the minimum of retweets needed to be classified as an idea starter (which are identifiable by being a red node). Yellow nodes indicate a user is a curator (a user that connects two more idea starters together), and blue nodes are amplifiers (those that were first to retweet a chain of retweets). Orange nodes however indicate users that are not only curators, but amplifiers as well, these are people that are first to retweet, and also are the links between idea starters. The remaining green nodes are commentators, the users who don't fit into any of the other groups. The scale of the Red and Orange or Blue nodes are an indication on the users ranking within their category, for instance, a large Red node would indicate a user with a high level of retweets, proportionally more than the minimum level set. A large Orange or Blue node indicates that the user has been an original retweeter proportionally more times than others

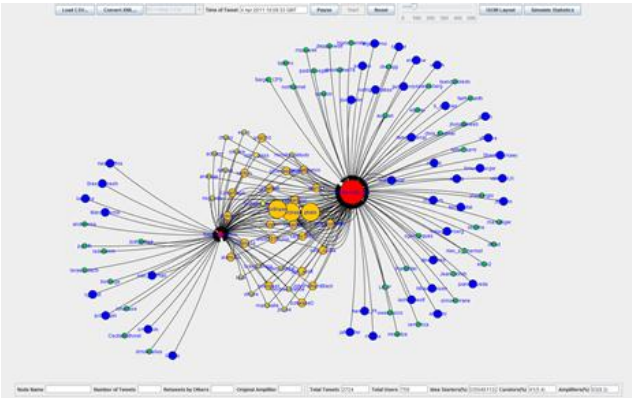

Figure 5 ReFluence – Visualization of Retweet Graph

The statistics simulation works by examining the number of classified influential users based on the RT^{Min} value set. The simulation works by starting at a value of 1, incrementing by 1 until there are no more users categorized as influential. This then gets plotted as a graph of percentage of users against RT^{Min} value.

Concluding Remarks

ReFluence is still an application in development and further work is being done on examining different methods to identify users based on their network behavior. The future aims for ReFluence is for it to become part of the *Web and Internet Science*[5] (WAIS) Web Science toolkit for observing and exploring the Web.

References

Bentwood, J. (2008). Distributed Influence: Quantifying the Impact of Social Media. *Edelman*. Retrieved March 1, 2011, from http://technobabble2dot0.files.wordpress.com/2008/01/edelman-white-paper-distributed-influence-quantifying-the-impact-of-social-media.pdf

Cha, M., & Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social*.

Chu, Z., Gianvecchio, S., & Wang, H. (2010). Who is Tweeting on Twitter: Human, Bot, or Cyborg? *Proceedings of the 26th Annual Computer Security Applications Conference* (pp. 21-30).

Hargreaves, J., & Davies, L. (2011). *New Influentials, Collective Emotional Intelligence & The " Topology of Influence." Science*.

Paul, M. J., & Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. *Artificial Intelligence*, 265-272.

Pear Analytics. (2009). *Twitter Study – August 2009*.

Twitter. (2011). Twitter Blog: #numbers. Retrieved October 1, 2011, from http://blog.twitter.com/2011/03/numbers.html

Welch, M. J., Schonfeld, U., & He, D. (2011). Topical Semantics of Twitter Links. *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 327-336).

Weng, J., & Lee, B.-sung. (2011). Event Detection in Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. AAAI.

Zhou, Z., & Kong, J. (2010). Information Resonance on Twitter : Watching Iran. *Network*, 123-131.

[3] https://dev.twitter.com/docs/rate-limiting

[4] http://jung.sourceforge.net/

[5] http://www.wais.ecs.soton.ac.uk/