Identifying User Roles in Twitter

Ramine Tinati
School of Electronics and Computer Science,
University of Southampton,
Southampton. England
rt506@ecs.soton.ac.uk

Leslie Carr
School of Electronics and Computer Science,
University of Southampton,
Southampton. England
lac@ecs.soton.ac.uk

ABSTRACT

Social Networks such as Twitter offer a platform for individuals to create and share messages, establish 'friendships' between each other, and even become part of specific communities. Twitter has enabled a range of important social activity to succeed, including identifying public health issues and more recently, as a platform for social and political change. However, in spite of this, the volumes of messages that are transmitted per day make identifying valuable content from the back chatter and ultimately, influential individuals from spam, difficult.

To tackle this, a classification model which utilizes the features offered in Twitter has been developed which classifies users based on their interaction behavior. This model helps identify Twitter users into specific categories based on their own specific behavior. This provides a method of identifying users who are potentially producers or distributers of valuable knowledge.

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory]:** Value of information

General Terms

Algorithms, Design, Experimentation, Measurement, Theory

Keywords

Twitter, User Classification, Influence, Web Science

1. INTRODUCTION

In recent years, the development of social networking technologies has proved to be one of the fastest growing activities on the Web – both in its development and usage; current available figures show that social networking sites like Facebook have over 800 million users [1], and Twitter with over 100 million active users [2].

Inevitably, with the gigantic usage brings gigantic steams of information; Twitter recently recorded over 250 million tweets a day. Although this data has proven to be useful for a number of different activities which provide benefit to society [3–5], based on a recent analysis of Twitter data, up to 40% of the messages passed can be classified as white noise [6].

Focusing on the Twitter service, the amount of the data available provokes the question of how can we identify the valuable information from the rest? There does exist various approaches to distil the information, including spam detection [7], [8], various forms of sentiment analysis [9–11] and also qualitative studies examining meaning behind tweets [12]. These approaches do offer a way to help identify valuable users based on their individual Twitter data streams (the Tweets); however we propose that another way to extract the valuable information can be found by examining the propagation of messages that flow between users.

This is made possible by Twitter's retweet feature, which enables users to republish someone else's tweet to their own timeline of tweets; and by doing so provides a back link to the original author, thus providing a traceable link between Twitter users and tweets. Although the concept of retweeting is fairly recent, there has been some qualitative research conducted on the reasons for retweeting [13], and also research indicating that tracing the retweets of Twitter users is a useful and appropriate metric to measure the importance of users within the network [14].

Based on the findings of the discussed research, a model was developed which utilized the Twitter's retweet functionality to help identify different users within a given network.

2. TWITTER USER CLASSIFICATION MODEL

The classification of the users is based on ongoing work with Edelman – a personal relations company interested in finding ways to obtain influence individuals in social networking technologies. Edelman's *Topology of Influence* [15], a user classification scheme based on their long established professional knowledge provided us with a starting point on how to categorize individuals based on their characteristics. This was then adopted to reflect the technical and social architecture of Twitter and its retweet functionality.

Three categories were chosen as representable user types on twitter: *idea starters*, users who have a large proportion of their tweets retweeted, thus suggesting their ideas are important and are of value to share. *Amplifiers*, users who are the first to spot an important tweet and first to retweet it, which eventually become part of retweet chain. *Curators*, users who spot multiple influential users on Twitter and retweet them, thus acting as an aggregator of valuable content.

Definition 1 Calculating an Idea Starter

$$\frac{\sum U^{rt}}{RT^{min}} > 1$$

Where U^{rt} is number of retweets of a user and RT^{min} is minimum retweet number

Definition 2 Calculating an Amplifier

$$\frac{\sum U^{t}}{\sum RT^{u} \times \sum rt^{orig}} > 1$$

Where U^{t} is number of user's tweets, RT^{u} is number of user's retweets, and RT^{orig} is number of retweets which were first in retweet chain

Definition 3 Calculating a Curator

$$\frac{\sum RT^{u}}{\sum U^{uniqRT}} > 2$$

Where RT^{u} is number of a user retweets and U^{uniqRT} is number of unique number of users that a user has retweeted.

3. IMPLEMENTATION

Based upon the Twitter data available (a timeline of tweets including data on Twitter usernames, tweet text, and timestamp) and the classification model provided by Edelman [15], the implementation aimed to produce a visual tool to examine the growth of a Twitter retweet network over a given time period. The low level model of the system architecture is shown in Figure 1; this enables nodes (users) and edges (retweets) to be constructed. This in combination with the rules stated in Definition 1 and 3 provides a way to identify *idea starters* and *curators*.

To model *amplifiers*, the growth of the retweets chains over a time period thus the propagation of a tweet requires modeling. As shown in Figure 2, based on the timestamp of the tweet, the original tweet can be found and then the chain of retweets can be constructed from this. This then can be used in conjunction with rules from Definition 2 to identify *amplifiers.*

4. CONCLUDING REMARKS

Through exploring a number of datasets, the model has demonstrated that the method of classifying different user types provides an alternative approach to identifying and extracting important and valuable Twitter data, both users and tweets.

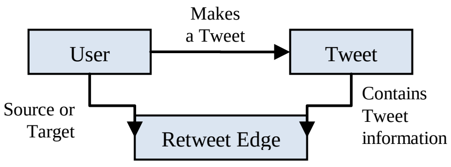

Figure 1 Data Model Overview

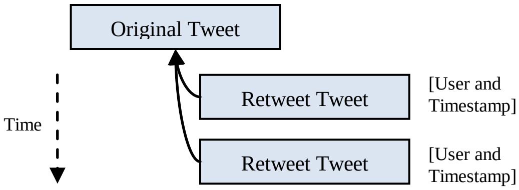

Figure 2 Retweet Chain

5. ACKNOWLEDGEMENT

This research was funded by the Research Councils UK Digital Economy Programme, Web Science Doctoral Training Centre, EP/G036926/1.

6. REFERENCES

[1] Facebook, "Facebook - Statistics," 2011. [Online]. Available: https://www.facebook.com/press/info.php?statistics. [Accessed: 01-Dec-2011].

[2] Twitter, "Twitter Blog: #numbers," 2011. [Online]. Available: http://blog.twitter.com/2011/03/numbers.html. [Accessed: 01-Oct-2011].

[3] J. Weng and B.-sung Lee, "Event Detection in Twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[4] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonc, A. Flammini, and F. Menczer, "Political Polarization on Twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 89-96.

[5] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," *Artificial Intelligence*, pp. 265-272, 2011.

[6] Pear Analytics, "Twitter Study – August 2009," 2009.

[7] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a twitter network," *First Monday*, vol. 15, no. 1, 2010.

[8] Z. Chu, S. Gianvecchio, and H. Wang, "Who is Tweeting on Twitter: Human, Bot, or Cyborg?," in *Proceedings of the 26th Annual Computer Security Applications Conference*, 2010, pp. 21-30.

[9] M. Pennacchiotti and A.-maria Popescu, "A Machine Learning Approach to Twitter User Classification," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 281-288.

[10] J. Weng, E.-peng Lim, and J. Jiang, "TwitterRank: Finding Topic-sensitive Influential Twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010.

[11] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. H. Chi, "Short and Tweet: Experiments on Recommending Content from Information Streams," in *Proceedings of the 28th international conference on Human factors in computing systems*, 2010, pp. 1185-1194.

[12] J. Hurlock and M. L. Wilson, "Searching Twitter: Separating the Tweet from the Chaff," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 161-168.

[13] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in *HICSS-43*, 2010.

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

[15] J. Hargreaves and L. Davies, "New Influentials, Collective Emotional Intelligence & The ' Topology of Influence'," 2011.