Reilly, C.F.; Salinas, D.; De Leon, D., "Ranking Users Based on Influence in a Directional Social Network," *Computational Science and Computational Intelligence (CSCI), 2014 International Conference on* , vol.2, no., pp.237,240, 10-13 March 2014

DOI: 10.1109/CSCI.2014.127

The published version of this paper is available from IEEE Explore:
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6822338&isnumber=6822285

Ranking Users Based on Influence in a Directional Social Network

Christine F. Reilly*, Dave Salinas, and David De Leon
Department of Computer Science
University of Texas – Pan American
Edinburg, Texas, USA
Email: reillycf@utpa.edu, dave.salinas@yahoo.com, dcdeleon@broncs.utpa.edu
*Contact Author

***Abstract*—This paper describes the preliminary work on a study that ranks users in a directional social network by their influence during a particular time period. Our method considers a user with high influence to be one who has a high ratio of forwarded messages to the number of messages she posts. After implementing our influence ranking program, we plan to compare our method with others in the literature. This comparison will evaluate the similarity of the lists of influential users that are generated by these different methods. We will also evaluate the ease of use and time required by each of the influence calculation methods.**

Keywords–big data in social media.

I. Introduction

The problem of identifying influential individuals in society has been scientifically studied since the 1890's [1]. However, until recently, theories about social influence have remained theories due to the difficulty of gathering data that could be used to validate these theories. The recent development of online social networks provides a rich source of data that can be used to study social influence. Identifying influential individuals is of interest to many parties, especially those who wish to have their message spread widely throughout the network. Political movements and marketing companies are examples of organizations who are interested in discovering influential individuals.

Online social networks are a widely used method for spreading information and news. Examples of online social networks that are currently popular include Facebook and Twitter. Facebook is an example of an undirected social network where the relationship between users must be reciprocal. In contrast, Twitter is a directed social network because the relationship between users need not be reciprocal. A study from 2010 found that only 21.1% of relationships in Twitter were reciprocal [2]. We use a directed social network in this work, and utilize messages and metadata from Twitter in order to test our methods.

We consider a user in a social network to be influential if the information she shares becomes widely disseminated throughout the network. It has been noted that forwarding of information is a narrow definition of influence [3]. A user may forward information based on their interest in the information itself, not due to the person who sent the information. Additionally, forwarding information may or may not indicate the presence of factors that influence purchasing behavior or political opinions. Despite these limitations, we proceed with examining the use of information passing as a measure of influence in social networks because that is the behavior that is easily observable in the network.

A. The Twitter Social Network

Twitter is a microblogging service where users post short messages of up to 140 characters. These messages are called tweets. The audience for a tweet is other Twitter users (called followers) who choose to subscribe to the tweet author's message feed. Unlike social networks such as Facebook, the following relationship in Twitter need not be reciprocal. Most users make their tweets public and any user may follow any other user without the consent of the followed user. By following a user, the follower receives the followed user's tweets.

Twitter users forward messages through a process called retweeting. We can illustrate the retweeting process by considering two hypothetical users, Alice and Bob. Bob follows Alice and reads her tweets. When Bob reads a tweet that he finds particularly interesting, he may choose to retweet that tweet. Now Alice's tweet will be sent to all of Bob's followers. Messages are spread through Twitter by cascades of retweets.

Because most users allow their tweets to be publicly broadcast, Twitter is an appealing social network to use for research purposes. There is a publicly available application programming interface (API) that we can use to write programs that gather messages and associated metadata from Twitter.

B. Measures of Influence in Twitter

Three ways of measuring influence in Twitter are indegree, retweets, and mentions. Indegree refers to the number of followers the user has. Retweets are the number of times a user's messages are forwarded in the network. Mentions are the number of times a user is referred to in another user's tweet.

Indegree measures the size of the audience of a user [4]. This could also be considered a measure of a user's popularity [2]. A user with many followers has the potential of a large audience for their tweets. We say that this is a potential audience, because the followers may or may not actually pay attention to the given user's messages.

Mentions are a good measure of how connected a user is to their immediate network [4]. Often, mentions are used to engage the mentioned user in a conversation. Other times, mentions are used when talking about the mentioned user. Therefore, a large number of mentions shows that a user has a high ability to engage in conversation.

Retweets measure how widely a user's message disseminates through the network [4], [2]. A highly retweeted message shows that the user has the ability to generate content that is valuable for other users to pass along to their followers.

This paper is organized as follows. Section II discusses related work. Our methods for calculating influence and collecting data are presented in Section III. We then discuss how we will count tweets and retweets, and compare our method with others in Section IV. Finally, in Section V we conclude the paper and discuss avenues for future work.

II. Related Work

A number of recent studies have focused on examining the properties of the Twitter social network, including calculating the influence of users.

Cha et. al. compared three measures of user influence: indegree, retweets, and mentions [4]. They ranked users based on each influence measure and used a correlation measure to determine how closely related a pair of influence measures are. They found that mentions and retweets were highly correlated, meaning that users with many retweets are mentioned often, and users with many mentions are retweeted often. Indegree had a low correlation with both retweets and mentions, indicating that the most connected users are not necessarily the most influential.

The above finding may seem counterintuitive. If a user has a large audience, shouldn't that mean that she is highly influential? The issue is that indegree measures the size of the audience for a particular user, but does not measure whether any of the audience members are actually listening. There are a few common practices in Twitter that may indicate why indegree is not a good measure of influence. First, some users follow a lot of other users but do not read all of those other users' messages. Second, some users will reciprocally follow everyone who follows them but, as in the first example, may not pay attention to all of the users they follow.

Kwak et al. also used the number of followers and the number of retweets to rank Twitter users [2]. Additionally, they ranked users with the PageRank algorithm, where the nodes were users and the directed edges represented one user following another. They found that number of followers and PageRank gave similar rankings to the users. We are not surprised by this finding because the PageRank method was based on the graph of followers. Ranking users by the number of retweets provided a very different ranking of users than the other two methods.

Another study that used information propagation in Twitter as a measure of a user's influence tracked the spread of bit.ly URL's (bit.ly is a URL shortening service) through the graph of who follows whom in Twitter [3]. They claim that this is a more inclusive measure of influence than using retweets because a user may see a URL in a message from someone they follow, and then include that URL in a new tweet without acknowledging the source user. Additionally, tracking URL's gives an upper bound measure of influence because a user may independently post the same URL as one of the users she follows. This study then examined what factors best predict future influence and found that individuals who have been influential in the past and have many followers are likely to be influential in the future. The authors of the study note that the messages posted by these influential individuals will not always be influential because they found few deep message cascades in Twitter.

III. Methods

Because prior work has indicated that the number of retweets may be a better measure of influence than the number of followers or other metrics, we focus our calculation of influence on retweets. We deviate from the prior work by focusing on the ratio of retweets per tweet, instead of on the sheer number of tweets as was done by Cha et al. [4] and Kwak et al. [2].

Our intuition is that the retweet to tweet ratio provides a better measure of overall influence. For example, a user could be a very prolific tweeter by posting a large number of messages. If each of these messages is retweeted once, then this prolific tweeter will have a high sheer number of retweets. However, her messages have not spread widely throughout the social network. Another example where the sheer number of retweets may not be a good measure of influence is the case where one of a user's tweets is widely forwarded through the network, but the rest of the user's tweets are not retweeted. In this case, the user does not generally have influential tweets, only the single tweet that was very popular. Using the retweet per tweet ratio as our measure of influence will be more likely to indicate that users who have many tweets that are highly retweeted are the most influential.

A. Calculating Influence

For a more formal definition of influence, we consider an influential user to be one who has a high retweet per tweet ratio during the given time period. Given the following terminology:

- I_i = Influence of user i in a certain time period
- M_i = The number of tweets written by user i in a certain time period
- N_i = The number of retweets of user i's tweets in a certain time period

Consider the influence of user i during a certain time period to be the ratio of retweets to tweets. Therefore, define the influence of user i during a certain time period as:

$$I_i = \frac{N_i}{M_i} \tag{1}$$

In order to rank users by influence, we normalize influence by dividing the influence of each user by the sum of the influences of all users. Given the terminology above, along with the following additional terminology:

- $\widetilde{I}_i$ = Normalized user influence
- U = total number of users

Define the normalized influence of user i during a certain time period as:

$$\widetilde{I}_i = \frac{I_i}{\sum_{k=1}^{U}(I_k)} \tag{2}$$

In Section IV we will discuss how we will use the normalized user influence to rank Twitter users.

B. Data Collection

In order to test our method for ranking Twitter users based on retweets, we need to count the tweets and retweets of those tweets for the currently active users. The first task is to identify the active users. Therefore, we use the Twitter Application Programming Interface (API) to gather information about current activity on Twitter, and store this data in a relational database. We are currently finalizing our data gathering programs and will begin data collection soon.

IV. Implementation

After collecting data from Twitter, we will be ready to run our influence ranking program. We will rank users by influence during a particular time period. As we will discuss in Section V, in the future, we plan to examine how the top influential users fluctuate over time.

During the specified time period, we will find all users who have posted an original message. Next we will find the number of times that each of the user's messages have been retweeted during the given time period. Then we can use the methods discussed in Section III-A to compute the influence of each user and then rank the users by influence.

A. Counting Retweets

In order to count retweets, we must first identify messages as being retweets. Twitter users utilize two different methods for retweeting: the native retweet and the organic retweet.

In late 2009, Twitter added native functionality for retweets to its API. This functionality adds a retweet button to each tweet, when viewed in Twitter's graphical user interface. When a user wishes to forward a tweet, she clicks the retweet button on that tweet to forward the message to her followers. The native retweet functionality gives retweets a uniform format and allows the API to provide methods for counting the number of retweets of a given tweet.

Prior to the addition of native retweet functionality, and still to this day, users employ a number of different organic retweeting practices [5]. These retweeting practices are called organic because they arose from within the Twitter community. The most common practice is to add "RT @username" to the beginning of the retweeted tweet, where "RT" indicates that the tweet is a retweet, and "@username" tags the user who is being retweeted. Other retweeting practices are used to indicate that the tweet has been significantly altered by the retweeter, or occurred simply because there was no standard protocol for retweeting.

Counting the number of native retweets is a simple matter of using Twitter's API. The API provides a method that returns a list of the users who retweeted a particular tweet. One limitation of this method is that we are unable to filter the retweeter list by time period. Our influence calculation method considers the number of retweets per tweet during a specific time period. We do not expect this to be problematic because previous studies of retweeting practices has shown that one half of retweets are posted within one hour, 75% within one day, and only 10% one month later [2].

In order to count organic retweets, we must scan through the tweets we collected in our database and count the number of times we find the "RT @username" string for the given user. Because our method of calculating influence counts retweets and tweets for a user, instead of counting the retweets for a particular tweet, we do not need to determine which of the user's tweets is being retweeted. All we need is the count of the number of organic retweets of that user during the specified time period.

As stated above, we simplify the identification of organic retweets by only searching for the "RT @username" syntax. This simplification means that we may miss some retweets that use alternative syntax for attributing the retweet. Additionally, we are not able to detect the fact that some users forward another user's message without attribution. The possibility that we may miss some organic retweets is the tradeoff that we must accept when working with real–world data.

B. Comparison with Other Methods of Ranking by Influence

In order to evaluate the usefulness of our method of ranking Twitter users by influence, we will compare our method with others that have been presented in the literature. We will compare the similarity of the ranked lists generated by the different methods as will as the ease of implementation of the methods.

We may find that different methods of ranking users by influence result in similar ranked lists of users. In this case, we will consider whether one method is easier to implement than the other. On the other hand, we might find that different methods generate very different rankings of influential users. This finding would require additional investigations to determine the causes of the difference.

The other ranking methods we will compare our method to are the ones that also consider retweets as a measure of influence [4], [2], [3]. We will compare the results my manually examining the top 20 users, and by using a mathematical comparison [4], [2].

V. Conclusions and Future Work

We have presented our preliminary work on ranking users in Twitter, a directional social network, by influence. Our methods are similar to prior work in that we also consider message forwarding as the measure of influence. However, we use the ratio of the number of forwards of a user's messages to the number of messages posted by the user, instead of the sheer number of times a user's messages have been forwarded.

This ranking of users in Twitter by influence sets the stage for a number of future studies. Related studies have computed influence over the time period of a month or two [4], [2], [3]. We are interested in examining influence over a much shorter time period, such as a day. We can then examine how the ranking of influential users changes over a relatively short period of time.

VI. Acknowledgements

This work was funded in part by the University of Texas–Pan American NSF ADVANCE grant.

References

[1] A. R. Pratkanis, "An invitation to social influence research," in *The Science of Social Influence: Advances and Future Progress*, A. R. Pratkanis, Ed. New York, New York, USA: Psychology Press, 2007, ch. 1, pp. 1–15.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW2010*, Raleigh, North Carolina, USA, Mar. 2010.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on Twitter," in *WSDM'11*, Hong Kong, China, Feb. 2011.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 10–17.

[5] d. boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *HICSS-43 IEEE: Kauai, Hawaii*, 2010.