Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi
Palo Alto Research Center, Inc.
Palo Alto, CA, U.S.A.
{suh, hong, pirolli, echi}@parc.com

Abstract— **Retweeting is the key mechanism for information diffusion in Twitter. It emerged as a simple yet powerful way of disseminating information in the Twitter social network. Even though a lot of information is shared in Twitter, little is known yet about how and why certain information spreads more widely than others. In this paper, we examine a number of features that might affect retweetability of tweets. We gathered content and contextual features from 74M tweets and used this data set to identify factors that are significantly associated with retweet rate. We also built a predictive retweet model. We found that, amongst content features, URLs and hashtags have strong relationships with retweetability. Amongst contextual features, the number of followers and followees as well as the age of the account seem to affect retweetability, while, interestingly, the number of past tweets does not predict retweetability of a user's tweet. We believe that this research would inform the design of sensemaking and analytics tools for social media streams.**

Keywords-Twitter; retweet; tweet; follower; social network; social media; factor analysis

I. INTRODUCTION

Among various microblogging systems, Twitter is the most popular service by far. Due to its ease for real-time information sharing, Twitter has impacted public discourse in the society. In Twitter, a lot of information is shared via its social network structure, but little is known about how and why certain information spreads more widely than others.

One interesting emergent behavior in Twitter is the practice of retweeting, which is the relaying of a tweet that has been written by another Twitter user. When a user finds an interesting tweet written by another user and wants to share it with her followers, she can retweet the tweet by copying the message. This can be done in one of two ways. First, one can retweet by preceding it with RT and addressing the original author with @. For example, "RT @userA: my experience with the new #iPad is great!" Second, Twitter also enables users to retweet easily with one-click.

As retweeting became an established convention inside Twitter, researchers have investigated retweeting as a conversational practice, such as how authorship, attribution, and communicative fidelity are negotiated in diverse ways [2]. Retweeting can be understood as a form of information diffusion since the original tweet is propagated to a new set of audiences, namely the followers of the retweeter. These retweeting actions are often associated with certain values of the original information items [2]. Retweeting may be to entertain a specific audience, to comment on someone's tweet, to publicly agree with someone, or to save tweets for future personal access. These actions suggest that the original tweet contains valuable information [2].

Another interesting investigation on retweeting is Zarrella's series of blog posts [12]. Zarrella showed that retweets have quite different content characteristics from normal tweets. For example, he reported that 56.7% of retweets have URLs in them while only 19.0% of regular tweets have URLs. This suggests that retweets are used to spread interesting web pages, videos, and other web content to other users. Zarrella's work focused mainly on direct content analysis of the retweets and the original tweets themselves, such as the most likely words to be retweeted, types of URL shortening services used, and reading grade level of the retweets. His recent posts have also examined the depth that tweets reach.

In our work, we are interested in extending his findings to understand factors that might affect retweetability of a tweet. We are motivated to investigate the feasibility of building a retweet model with simple measures. We are interested in not only just content features, but also contextual features. Content features that we examined include whether the tweet contains URLs, hashtags, and mentions (referencing other users in tweet text). Contextual features include the number of followers (people who follow me) and followees (people who I follow), the age of the account, the number of favorited tweets, and the number and frequency of tweets. Zarrella [12] examined a few of these features (namely URLs and followers), but not in detail.

We document in detail our analytical method and the way we collected our data set. We collected a significant number (74 millions) of tweets through the Twitter API, and studied various content and contextual features that have close relationships with retweetability of tweets. We quantitatively identified features that are significantly associated with retweet rate and built a predictive retweet model.

The rest of this paper is organized as follows. First, we discuss prior work on Twitter, focusing on retweeting practice. Next, we introduce the data sets used in the study, followed by explaining the features that might affect retweet rate of a tweet. We then describe our retweet model based on a GLM

(Generalized Linear Model) analysis. For interesting features (e.g. URL, hashtag, the number of followers), we provide further detailed analysis. We conclude with a discussion of the implications of our findings.

II. MICROBLOGGING AND TWITTER

Microblogging is a form of blogging in which entries typically consists of short content such as phrases, quick comments, images, or links to videos. Notable services include Twitter, Tumblr, Jaiku, Posterous, and Google Buzz. Recently, as microblogging services have gained wide popularity, users have adopted them for novel purposes including sharing news, promoting political views, marketing, and tracking real time events [2][7][11][13].

In addition to consumer usage, startups have begun to investigate how to adapt microblogging to enterprise environments (with example companies such as SocialCast, Jive, Yammer, etc). Both Zhao et al. [13] and Convertino et al. [4] investigated how to adapt microblogging to enterprise environments, suggesting that it can be tailored to facilitate informal communication between colleagues in organizations.

Among various microblogging systems, Twitter is the most popular service by far. Twitter is a social networking and microblogging service that allows users to send and read 140-character short messages known as tweets, enabling users to share and discover topics of interest in real-time. Users choose to follow other notable users to gain updates on news and statuses. Once authored by a user, tweets are immediately delivered to the author's subscribers or followers. For a reader, tweets from all users whom she follows are gathered together and displayed in a single reverse chronological list for consumption. Twitter also provides a set of application programming interface (http://apiwiki.twitter.com/), which allows third party applications to send and receive tweets.

Since its creation in 2006, Twitter has gained notability and popularity worldwide. As of March 2010, it had about 105M registered users and over 50M tweets were sent daily[4]. Kwak et al. [8] conducted a large-scale study to analyze the topological characteristics of Twitter and its power as a new medium of information sharing. From Twitter's public timeline Java et al. [7] collected 1,348,543 tweets created by 76,177 users, examining the topological and geographical properties of Twitter's social network. They identified a number of usage categories such as daily chatter, conversations, sharing information/URLs, and reporting news.

Perhaps the most popular usage is for users to inform others and to express themselves. Naaman et al. [9] examined the content of 3379 tweets by manually coding the messages collected from the public timeline, finding that 80% of the 350 users they studied posted messages relating to themselves or their thoughts, as opposed to sharing general news. Twitter also has been used politically by candidates in political campaigns. After the 2009 Iranian election, protesters used Twitter as a rallying tool and as a method of communication with the outside world.

Researchers have also examined how to build tools on top of Twitter. For example, Jansen et al. [6] investigated how Twitter is used to share consumer opinions about brands. Ramage et al. [10] and Bernstein et al. [1] proposed tools to group tweets into topics to support fast browsing. Chen et al. examined the personal recommendation of URL items in the Twitter stream [3].

Tweets are often used to converse with individuals or groups [5][7]. When a user wants to specify another user in a tweet, she can use the form of *mentioning* '@username', which is subsequently parsed and translated into a clickable hyperlink to the mentioned user. These user links enable the discovery of other interesting persons to follow and often facilitate a conversation. Furthermore, for a user, tweets containing that user's name will appear in a special "replies tab" (accessible at http://twitter.com/replies for logged-in users) notifying that the tweet was intended for her. If more than one person are included in a tweet using the @username format, each person will see the update in her own replies tab.

III. RETWEETS

Retweet is one particular case of mentioning. When a user finds an interesting tweet written by another Twitter user and wants to share it with her followers, she can retweet the tweet by copying the message, typically adding a text indicator (e.g. RT, Via) followed by the user name of the original author in @username format.

As discussed before, retweeting is associated with various social motivations such as entertaining a specific audience, commenting on someone's tweet, or publicly agreeing with someone [2]. Users often add more content or slightly modify the original when retweeting. Twitter users created a number of different conventions to retweet such as "RT @" and "via @" [2]. Retweeting has become so widespread that Twitter added a feature in 2009 to allow users to retweet easily with one-click.

Retweeting has become the key mechanism for spreading information in Twitter. Therefore, it is important to explore how retweet works to understand how information is diffused in the Twitter network. Building the retweet model might lead to the optimization of information diffusion that naturally occurs in the Twitter network. Specifically, we are interested in features that might affect the retweetability of tweets because we seek to explain why certain tweets spread more widely than others.

IV. DATA SETS

We collected two data sets for analyses in this paper. First, we selected 10,000 tweets (10K data set) and traced, as accurately as possible, the retweet count for each tweet. The 10K retweet dataset was used to perform an exploratory data analysis using Principal Components Analysis (PCA) and Generalized Linear Modeling (GLM).

In addition to the 10K data set, we also collected 74 million tweets (74M data set) using the Twitter open API. We used the 74M data set for quantitative content analysis associated with retweeting.

A. 10K Data Set

The purpose of our exploratory data analysis is to understand the features that are associated with retweeting.

Conceptually, one would like to focus on a set of tweets, use their features as independent variables, and treat the retweet rate as the dependent variable to be predicted from the samples.

For our exploratory data analysis, we arbitrarily chose a date and collected tweets on that date (1,772,906 random sample of all public tweets posted on March 1, 2010). Retweets were filtered from this sample to yield 1,560,217 original tweets. From this sample, we randomly selected 10,000 tweets. On March 19, 2010, we used Twitter's API[2] to determine the number of retweets for this 10K sample. Out of the 10K tweets, 219 tweets had been retweeted at least once and none of them had been retweeted more than 20 times. In addition, 344 of those 10K tweets had been deleted by users from Twitter later. This yielded a set of 9,656 tweets, which we used to perform our exploratory data analysis.

B. 74M Data Set

For the purposes of further exploration of the relationship of social links and content popularity to retweet rates, we used Twitter's streaming API[3] to collect a random sample of the public tweets from January 18, 2010 to March 08, 2010, yielding 73,884,474 tweets (approximately 1.5M tweets per day). Twitter has more than 50M tweets daily[4], so we estimate that the data set represents about 2~3% of all tweets.

C. Tweet Features

For data sets collected above, we extracted a set of features (Table I). We selected features that are language independent, due to our desire to build a retweet model that is universal across languages.

The first set of features is concerned with the content of the tweets (URL, hashtag, mention), and the second set is generally about the tweet authors (follower, followee, favorite, day, status).

TABLE I. TWEET FEATURES

URL	# of URLs in a tweet
Hashtag	# of hashtags in a tweet
Mention	# of usernames specified in a tweet excluding ones used for making a retweet (e.g. via @username, RT: @username)
Follower	# of users who follows the author of a tweet
Followee (Friend)	# of friends that the author is following
Days	# of days since the author created Twitter account
Status	# of tweets made by the author since the creation of the account
Favorite	# of favorited tweets by a user
Retweet	# of retweets recorded for a given tweet

The first eight features in Table 1 were extracted for all the tweets in the two data sets – 10K and 74M data set. The content features were extracted using regular expressions from text of tweets. The contextual features were directly acquired through calls to the Twitter API.

Unlike other features, we found that "Retweet" is tricky to acquire directly because (1) there is no generally accepted agreement as to what constitutes a retweet and (2) finding all subsequent retweets for an original tweet is not always possible since we only have 2~3% sample of all the tweets. For our analyses, we identified retweets in two different ways:

- *Regular Expression Method*: For questions that did not require the linkage of a retweet to the original tweet, we used a set of textual markers (as suggested in [2]). For example, to identify the number of retweets containing URLs, we scanned for text markers such as "RT @", "RT:@", "retweeting @", "retweet @", "via @", "thx @", "HT @", and "r @". Using such marker methods, we found 8,235,837 retweets in the 74M data set (11.15%). This set of 8.24M retweets was used to study the content features (i.e., URL, hashtag, and mention).
- *Feature Retweet Method:* The special one-click retweet feature in Twitter allows users to retweet with one-click[5]. These "feature retweets" are part of the 8.24M retweets identified by the regular expression method described above, but Twitter additionally provides some contextual information for these feature retweets through API calls. This method yielded 2,993,303 feature retweets (36.34% of the 8.24M retweets). The feature retweet data provide us with rich contextual information for the retweet as well as linkage back to the original tweet.

V. EXPLORATORY PRINCIPAL COMPONENTS ANALYSIS

We performed a Principal Component Analysis (PCA) with the 10K data set with the nine features. PCA is a data reduction technique in which possibly correlated features (e.g., those in Table I) are transformed into a smaller number of factors called principal components. The technique is often used in an attempt to reveal internal (underlying) structure that maximally accounts for the variance in the data set. Our PCA analysis used the varimax rotation technique to produce orthogonal factors. Table II presents factors (principal components) extracted by PCA. The eigenvalues (the second column in Table II) represent the variance accounted for by each factor, which are also presented as a percentage of the total variance accounted for (the third column in Table II). The factors are extracted in descending order of variance accounted for in the original set of features. The last column of Table II presents the cumulative total variance accounted for with the addition of each successive factor.

Two rule-of-thumb methods are usually proposed for identifying the "right" number of factors to retain to represent the data variance. The Kaiser Criterion recommends retaining all factors with eigenvalues greater than 1. The Scree plot test recommends plotting the percent variance as a function of the factor number and choosing factors that occur before a flattening in the slope. Together these rules suggest retaining Factors 1, 2, and 3 in Table II, which account for 44.34% of the total variance.

[2] http://apiwiki.twitter.com/Twitter-REST-API-Method:-A-GET-statuses-id-retweeted_by

[3] http://apiwiki.twitter.com/Streaming-API-Documentation

[4] http://blog.twitter.com/2010/02/measuring-tweets.html

[5] http://blog.twitter.com/2009/11/retweet-limited-rollout.html

Table III shows factor loadings for the features in Table I against the factors 1, 2, and 3 identified in Table II. Figures 1 and 2 are factor maps of the features in Table III. These factor maps summarize the (unrotated) factor loadings (correlations) of the original features in Table I with each of the three factors in Table III.

TABLE II. FACTORS (PRINCIPLE COMPONENTS) FROM THE ANALYSIS

Factor	Eigenvalue	% Variance	Cumulative % Variance
1	1.52	16.94	16.94
2	1.31	14.52	31.46
3	1.16	12.88	44.34
4	0.98	10.90	55.25
5	0.94	10.48	65.73
6	0.89	9.94	75.67
7	0.78	8.70	84.37
8	0.71	7.89	92.26
9	0.70	7.74	100.00

TABLE III. FACTOR LOADINGS

Tweet Features	Factor 1	Factor 2	Factor 3
Retweet	0.2870	-0.0921	0.4459
Hashtag	0.1709	-0.4035	-0.1355
Mention	-0.1555	0.6569	0.2732
URL	0.3422	-0.6417	-0.1883
Days	0.3868	0.3117	-0.0795
Favorite	0.3030	0.3247	-0.5632
Follower	0.5605	0.0325	0.5064
Followee	0.6382	0.0237	0.2524
Status	0.5561	0.2964	-0.4343

In Figure 1 and 2, each feature in Table I is mapped into a vector in the graphs to represent its correlation with factors, which denoted as axis in the graph. For example, the URL feature is represented as a vector pointing (0.3422, -0.6471) in Figure 1 according to its respective factor loading to Factor 1 and Factor 2 in Table III. Similarly, the URL feature is represented as a vector (-0.6417, -0.1883) in Figure 2 to represent the feature correlations with Factor 2 and Factor 3.

A. Interpretation of the Factors

Two of the factors (Factor 1 and Factor 3) appear to distinguish tweets by the profiles of the tweet authors. Factor 2 distinguishes tweets based on their contents. Inspection of Figures 1, 2 and Table III shows that retweets are correlated with Factors 1 and 3.

We interpret Factor 1 (Figure 1) as capturing the degree to which tweet authors are "broadcasters". Factor 1 is correlated with number of followees, number of followers, and number of tweets (status). Retweets correlate 0.29 with Factor 1.

We interpret Factor 2 (Figures 1 and 2) as a content factor separating tweets that contain URLs and hashtags from those with mentions instead. URLs and hashtags correlate positively with Factor 2, whereas mentions have a negative correlation. This makes sense given that tweets are limited in the amount of content they can communicate, so having one kind of content (e.g., URLs) will tend to be exclusive of another (e.g., mentions).

We interpret Factor 3 (Figure 2) as distinguishing types of different users, specifically separating tweet authors who get retweeted frequently and have lots of followers, from those who tweet frequently (status) and have many favorites. Factor 3 is positively associated with retweeting and number of followers, and negatively associated with status and favorites. One example that fits this category would be Bill Gates (@billgates). He made only 128 tweets so far and has no favorite items but he has more then 780 thousand followers.

In summary, Factor 1 is strongly associated with features representing the author profile: numbers of followees, number of followers, and status all load heavily. Factor 2 is associated with content features, showing strong correlations with mentions and negative correlations with URLs and hashtags. Factor 3 is strongly associated with retweets correlated with high number of followers but negatively correlated with number of favorites and status.

B. Structure of Features Associated with Retweeting

Examination of the retweet vector in Figures 1 and 2 reveals what other vectors appear to aim strongly in the same (or opposite) direction, which suggests positive (or negative) correlations among the features. Figure 1 suggests that retweeting is associated with the features of numbers of followees and followers as well as the content features of URLs and hashtags. Figure 2 additionally suggests that retweets are negatively associated with number of tweets (status) and number of favorites.

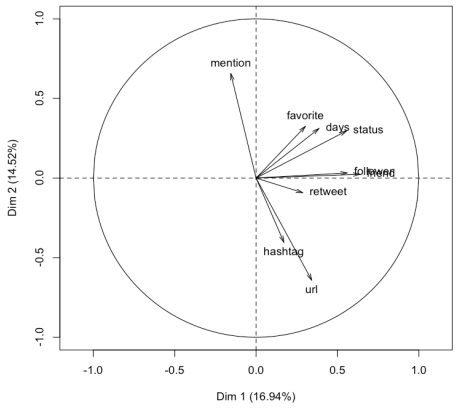

Figure 1. PCA factor map with Factor 1 (x-axis) and Factor 2 (y-axis)

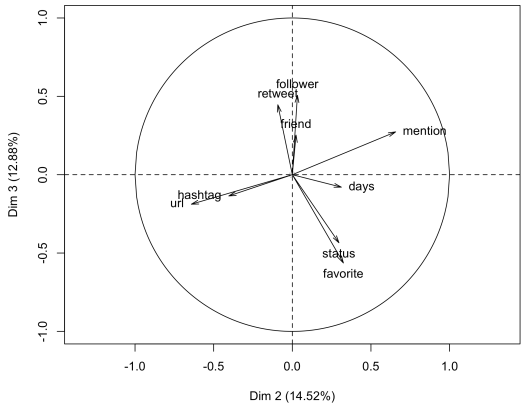

Figure 2. PCA factor map with Factor 2 (x-axis) and Factor 3 (y-axis)

VI. GENERALIZED LINEAR MODEL

The exploratory data analysis with PCA revealed underlying associations of retweeting with content and contextual features. We also wanted to directly capture the degree to which the probability of retweeting can be predicted from the first eight features in Table I. Towards this end, we fit a Generalized Linear Model (GLM) to the 10K data set. This results in a set of predictor coefficients for each feature that can be used in a logistic equation to predict the probability of a retweet. A GLM model is also equivalent to a predictive connectionist model (a single-layer perception with logistic thresholding).

TABLE IV. GENEARALIZED LINEAR MODEL

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.42000	0.146400	-30.19	**0.0000***
Days	0.00122	0.000296	4.12	**0.0000***
HashtagOrNot	1.32800	0.160300	8.28	**0.0000***
MentionOrNot	-0.29490	0.166800	-1.77	0.0771
URLOrNot	0.76360	0.150900	5.06	**0.0000***
Followee	0.00006	0.000020	2.85	**0.0043***
Follower	0.00002	0.000005	3.82	**0.0001***
Status	-0.00002	0.000009	-1.71	0.0876
Favorite	-0.00004	0.000163	-0.26	0.7987

Table IV presents the results of the GLM analysis. Several of the features were transformed into binary variables (0,1) and have the suffix "OrNot." The GLM corroborates the finding that the content features of hashtags and URLs have significant effects of retweet probability. Mentions have a marginally significant negative association with retweeting. The GLM also indicates that author features of number of followees and followers are strongly predictive of retweet probability. Number of status is marginally negatively associated with retweeting. Number of favorites is not a significant predictor of retweeting.

VII. TWEET FEATURES AND RETWEET RATE

The GLM model provides us a general picture about the correlation between retweeting and tweet features. Given the finding that some features have strong relationship associated with the retweetability of tweets, we further investigate relevant feature patterns in the 74M date set.

A. URL

We studied the impact of having a URL in the tweet. We searched for tweets and retweets containing URLs in the 74M data set. The result shows that 21.1% of tweets have at least one URL in their text. On the other hand, when we examine retweets only, we find that 28.4% have URLs in them. This finding matches the result from the GLM model. That is, a tweet with URLs is more likely to get retweeted.

Interestingly, our result differs somewhat from Zarrella [12], who found that 18.96% of tweets contain an URL and 56.69% of retweets include an URL. Further research is required to understand what contributes to this discrepancy.

In addition to the global pattern, we also investigated if there is a difference in retweet rate depending on types of URL. To do that, we calculate retweet rate for top domains.

One notable challenge to perform this domain-based analysis is the necessity of un-shortening URLs. In Twitter, due to the 140-character limit, it is a common practice to use a URL shortening service (e.g. http://bit.ly) when including URLs in tweets. Since we are interested in retweetability per domain, we perform unshortening to identify the original URL for each shortened URL.

TABLE V. RETWEET RATE FOR TOP 10 MOST POPULAR DOMAINS IN TWITTER WITH SOME OTHER NOTABLE SITES

Rank	Domain	In Tweet	In Retweet	Retweet Rate
1	http://twitpic.com	793680	129692	1.47
2	http://myloc.me	533082	121950	2.05
3	http://www.facebook.com	481349	55186	1.03
4	http://www.youtube.com	475509	79404	1.50
5	http://formspring.me	455377	2566	0.05
6	http://www.twitlonger.com	349760	236435	6.06
7	http://tweetphoto.com	258049	49676	1.73
8	http://youtu.be	196557	7508	0.34
9	http://twitcam.com	159684	2187	0.12
10	http://twitter.com	144002	39127	2.44
14	http://foursquare.com	90328	1763	0.18
19	http://www.flickr.com	47181	7599	1.44
20	http://mashable.com	43722	17778	3.65
25	http://news.bbc.co.uk	36286	6103	1.51
29	http://www.nytimes.com	31339	9035	2.59

With the unshortened URLs, we calculate a tweet rate per domain. For example, http://www.youtube.com is the 4th most popular domain in our data set. Among 74 million tweets, 475,509 tweets have URLs from that domain (e.g. http://www.youtube.com/v/xyz1234). We found that 79,404 retweets contain the same domain.

We first compute the retweet rate for the domain as the retweet number divided by the tweet number. We then normalized the rate so that a value of 1.0 represents the average retweet rate on tweets.

For this youtube.com example, the retweet rate of 1.50 (4th rank in Table V) is calculated by 79,404/475,509 * *Norm*. The normalization factor *Norm* is 74M (total tweets) / 8.2M (total number of retweets identified by the *Regular Expression Method*). When the retweet rate of a certain URL is 1.0, tweets having the URL are retweeted at the same rate as any other tweets would be. However, if the retweet rate of a URL is 2.0, tweets with the specific URL is two times more likely to be retweeted.

As shown in Table V, the retweet rates vary depending on URL domains. For example, formspring.me, which is the 5th most popular domain in Twitter, shows only 0.05 retweet rate hinting that tweets having that URL domain are very unlikely to get retweeted. On the other hand, the retweet rate of twitlonger.com is 6.06, which suggests that tweets with that domain in them have high retweetability.

To further understand the relationship between the popularity of domain and the retweet rate, we create a graph of the retweet rate for top 50 domains as shown in Figure 3. Each

data point represents a domain. X-axis is the popularity rank in Twitter based on how often tweets contain URLs of that domain in the 74M data set. Y-axis is the retweet rate as explained above. As seen in Figure 3, not all popular domains are also popular in retweets. The result shows that not all domains have the same retweetability.

Overall, the analysis showed that URL is a significant factor impacting retweetability and the domain of URLs also matters.

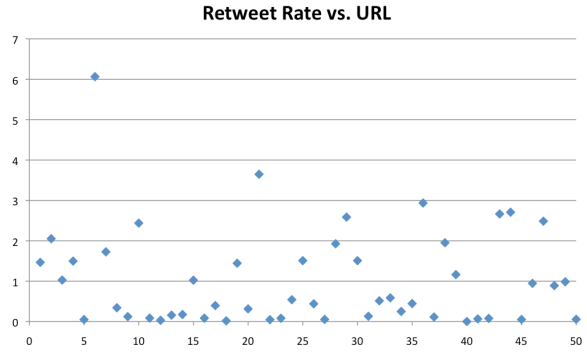

Figure 3. Retweet rate for top 50 most popular domains appearing in tweets (x-axis: popularity rank of domain in tweets, y-axis: retweet rate of domain)

B. Hashtag

The hashtag feature is represented by the number of hashtags included in the tweet text. Hashtag is a simple freeform keyword preceded by a character "#" (e.g. #nowplaying). In Twitter, a hashtag is translated into a clickable link that facilitates an easy search of tweets having the same hashtag. Hashtags are frequently used to represent topical keyword.

We searched for hashtags in tweets and retweets in the 74M data set. The result shows that 10.1% of tweets have at least one hashtag in their text while 20.8% of retweets contains hashtag. This observation matches our retweet model that a tweet with hashtags is more likely to get retweeted.

TABLE VI. RETWEET RATE FOR TOP 10 MOST POPULAR HASHTAGS

Rank	Hashtag	In Tweet	In Retweet	Retweet Rate
1	#nowplaying	355147	29846	0.75
2	#ff	224760	62331	2.49
3	#jobs	124728	2173	0.16
4	#fb	87959	10994	1.12
5	#tinychat	67225	273	0.04
6	#vouconfessarque	51578	43628	7.59
7	#fail	49248	9759	1.78
8	#tcot	47394	18527	3.51
9	#1	47373	9124	1.73
10	#followfriday	39986	11170	2.51

We wanted to further investigate relevant patterns in the 74M data set. By using the same analysis method as used for the URL feature, we computed retweetability for hashtags (Table VI). For each hashtag, we count the number of tweets and retweets having the hashtag, respectively. The retweet rate is calculated as a normalized ratio of the number of retweets to the number of tweets.

In Figure 4, each data point in the graph represents an individual hashtag. X-axis is the popularity rank of hashtags in Twitter based on how many times each hashtag appear in the 74M data set. Y-axis denotes the retweet rate of hashtags as the same way as in the URL analysis. The graph shows a very similar pattern to that with the URL feature case. Not all popular hashtags in tweets are popular in retweets. The result suggests that type of hashtag also does matter for the retweetablity of tweets.

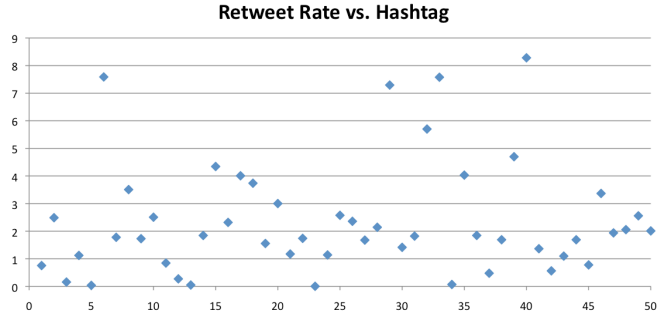

Figure 4. Retweet rate for top 50 most popular hash tags (x-axis: popurarity rank of hashtag in tweets, y-axis: retweet rate of hashtag)

C. Follower and Followee

In Twitter, *Followers* denote people who follow a user and *followees* represent "friends" whom a user follows.

Earlier, the GLM analysis showed that these two features have a very strong relationship with retweet rate. We further examined the relationships using the 74M data set. To do this, we calculated the retweet rate of users with different number of followers and followees.

We first analyzed users with differing number of followers and possible relationship to retweet rate. In Figure 5, on the X-axis, we put users into buckets according to an interval of around 100 followers, ranging from 0 to roughly 5000. On the Y-axis, we plot the retweet rate of users in that particular bucket.

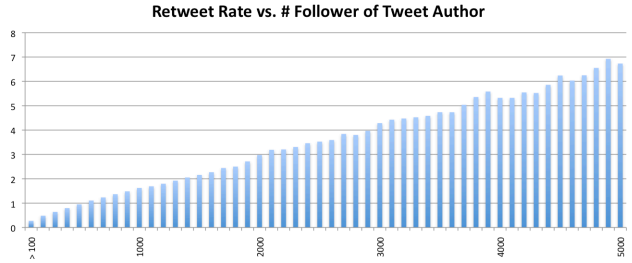

Figure 5. Historgram: Retweet rate (y-axis) and the number of followers of the tweet author (x-axis)

Again, the retweet rate represents a normalized ratio of the number of retweets to the number of tweets. For example, the leftmost bar in Figure 5 denotes a retweet rate of 0.27 from tweets that comes from Twitter users that have 0~99 followers. The number is calculated as follows. Among the total of 74M tweets, Twitter users that have less than 100 followers authored 34.9M tweets. Then, out of the 2.99M retweets identified by

the feature method (see section IV), we counted retweets of which the original author has less than 100 followers. We found 0.39M such cases. We calculate a (not yet normalized) retweet rate as a ratio of the number of retweets to the number of tweets. We then normalized the rate so that a value of 1.0 represents the average retweet rate. The normalization factor *Norm* here is 74M (total tweets) / 2.99M (total retweets identified by the *Feature Retweet Method*). In this example, a retweet rate 0.27 is calculated by 0.39M/34.9M * *Norm*.

As shown, Figure 5 shows a very strong linear relationship between the number of followers (x-axis) and retweet rate (y-axis). In other words, intuitively, the larger is the audience, the more likely the tweet gets retweeted.

Figure 6 show a correlated relationship between the number of followees of a tweet author and retweet rate. The larger number of *followees* a Twitter user follows, tweets from her are more likely to be retweeted. However, as shown in the figure, the relationship is not as strong as that with *follower*. This result suggests that the more sources the user follows, the more interesting the user's tweets are, probably due to the diversity of opinion and information items consumed by the user.

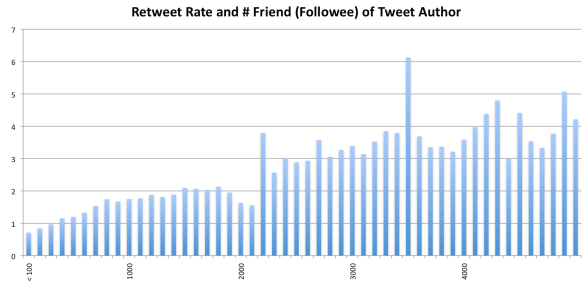

Figure 6. Histogram: Retweet rate (y-axis) and the number of followees of the tweet author (x-axis)

D. Past Tweets (Status)

The *status* feature is represented by the total number of tweets posted by a Twitter user since her account was created. Surprisingly, this feature does not have a strong relationship with retweet rate. As shown in Figure 7, retweet rates do not show any obvious pattern. When studied carefully, the front of the distribution seems to show some correlation but other than that we were not able to find a meaningful pattern. This observation matches the finding in the earlier GLM analysis.

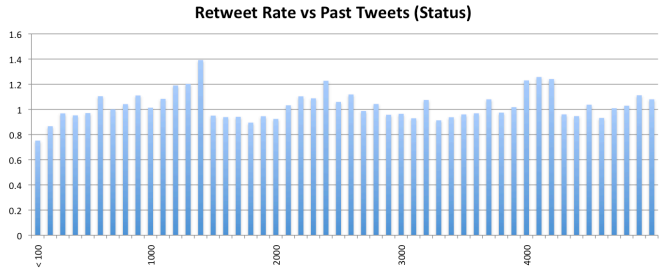

Figure 7. Histogram: Retweet rate (y-axis) and the number of past tweets made by the tweet author (x-axis)

In addition to the total number of past tweets, we also investigated the relationship of the retweet rate with the average number of daily tweets, which is calculated by the number of total tweets divided by the number of days since the author of a tweet created the account in Twitter. We do not include the details here for brevity but the analysis result shows that there is no significant relationship between the average number of daily tweets and the retweet rate.

E. Age of Twitter Account (Days)

Figure 8 shows varying retweet rates depending on the age of Twitter accounts. As shown in the graph, the seniority of Twitter users has a significant relationship with retweet rate. Tweets made by Twitter users who created their accounts more than 300 days ago shows a retweet rate higher than the average. While tweets authored by junior Twitter users exhibit low retweet rate in general, it is interesting to observe the increased retweet rate of Twitter users who created their accounts very recently (< 30 days), resulting in the slight U-shaped curve as shown in Figure 8.

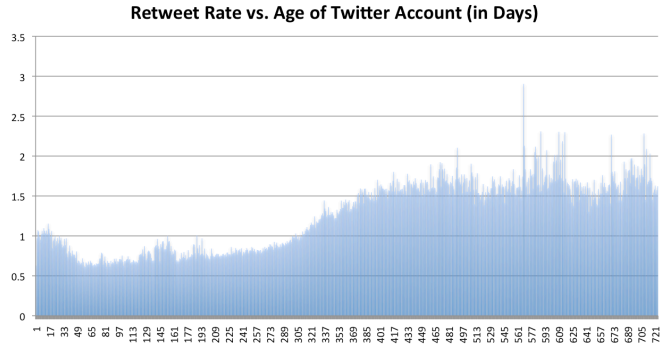

Figure 8. Histrogram: Retweet rate (y-axis) and the days since the author created the account (x-axis)

F. Favorites

The favorite feature is the number of tweets that are favorited by a Twitter user. Our study shows that the *favorite* feature is not heavily used and provides little benefit to understand the retweeting practice in Twitter. In the 74M data set, we found that 42.5% of tweets are made by users with no favorited items. The result shows that only 7.2% of tweets are coming from Twitter users with more than 100 favorite items (76% of users has less than 10 favorite items) and 8.7% of retweets are made by authors with more than 100 favorited items.

VIII. Discussion

In this paper, we have investigated the relationships between the tweet features and retweetability. We found that, amongst content features, URLs and hashtags have strong relationships with retweetability. Amongst contextual features, the number of followers and followees as well as the age of the account seems to affect retweetability, while, interestingly, the number of past tweets does not predict retweetability of a user's tweet.

One limitation of this study is the use of sampled tweets. Due to this issue, we were not able to collect all subsequent retweets for a given tweet – not all retweets are available in the dataset. Instead, for some analysis, we used only "*Feature Retweet*" to represent retweets in Twitter. Further research is required to investigate the validity of our method.

One interesting finding during our analysis of the URL domains was that the retweet rate vary depending on the types of domain. For example, we were able to observe that personal media domains such as justin.tv and twitcam.com have very low retweet rate (< 0.15) while trivia sites such as omg-facts.com and real-time discovery engines such as holykaw.alltop.com have high retweet rate (> 4.0). Among news media sites, retweet rates also vary depending on their sub-types. Tweets with URLs from some news sites such as mashable.com, theonion.com, and nytimes.com showed a high retweet rate (> 2.5) while tweets containing news.yahoo.com and news.google.com displayed a low retweet rate (< 0.6). Further work is required to understand the dynamics between contents of the URLs and retweet rate.

We also found that the number of past tweets does not seem to correlate with the probability of being retweeted. This suggests that broadcasting more often to your audience in Twitter does not necessarily lead to greater engagement.

Overall, the above findings suggest that retweetability has a very close relationship with social network context of the authors and the informational content and value contained in tweets.

In this paper, we are motivated to investigate the feasibility of building retweet model with easily computed features, because of our desire to build a recommendation engine for personalized Twitter streams. We want to enable users to catch interesting items that she might have missed in large number of tweets [3]. The hypothesis is that recommendations and personalization could help to optimize the information diffusion that already naturally occurs in the Twitter network.

The above findings suggest the importance of using social context in building recommendation engines for information streams like Twitter. Due to its short content, not all traditional recommendation and search algorithms work well on tweets. One interesting characteristic of microblogging is that it comes with a social context. For example, even a short message in Twitter has rich meta-information about its author as well as who might be listening, and have followed up. Such social context can be useful to determine the value of information. In this paper, we briefly touched the topic by including the number of followers and followees in the model. We believe that using such social context would play an increasingly important role in designing search, summarization, and recommendation tools for social media [1][3].

IX. Conclusions

Retweeting is the key mechanism for information diffusion in Twitter. We believe it is important to explore how retweet works to understand how information is diffused in Twitter network. To understand why certain tweets spread more widely than others, we investigated a number of tweet features that have potential relationship with the retweetability of tweets. In this paper, we quantitatively identified factors that are significantly associated with retweeting. We also built a predictive retweet model using Generalized Linear Model, and discussed the features of the model.

Generally, we found that, amongst content features, URLs and hashtags correlate with retweetability. Amongst contextual features, the number of followers and followees as well as the age of the account seem to affect retweetability, while, interestingly, the number of past tweets does not predict retweetability of a user's tweet. Overall, we hope that this research would inform the design of sensemaking and analytics tools for Twitter streams as well as other social information streams.

Acknowledgement

This research was sponsored in part by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

[1] Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., Chi, E. H. Eddi: Interactive Topic-based Browsing of Social Status Streams. To appear in ACM User Interface Software and Technology (UIST) conference, 2010.

[2] boyd, d., Golder, S., and Lotan, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. Proc HICSS'10, 1-10.

[3] Chen, J., Nairn, R., Nelson, L., Bernstein, M, and Chi, E.H. Short and Tweet: Experiments on Recommending Content from Information Streams. Proc CHI'2010, 1185-1194.

[4] Convertino, G., Kairam, S., Hong, L., Suh, B., and Chi, E.H. Designing A Cross-channel Information Management Tool for Workers in Enterprise Task Forces. In Proc AVI'10.

[5] Honeycutt, C. and Herring, S. Beyond Microblogging: Conversation and Collaboration via Twitter. Proc HICSS'09, 1-10.

[6] Jansen, B. J., Zhang, M., Sobel, K., and Chowdhury, A. Twitter Power: Tweets as Electronic Word of Mouth. Journal of American Society for Information Science & Technology, 60(11), 2169-2188.

[7] Java, A., Song, X., Finin, T., and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. Proc WebKDD/SNA-KDD'07, 56-65.

[8] Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a Social Network or a News Media? Proc WWW'10, 591-600.

[9] Naaman, M., Boase, J., and Lai, C.H. Is it Really About Me? Message Content in Social Awareness Streams. Proc CSCW'10, 189-192.

[10] Ramage, D., Dumais, S. T., and Liebling, D. Characterizing Microblogging Using Latent Topic Models. In Proc ICWSM'10.

[11] Wright, A. Mining the Web for Feelings, Not Facts. New York Times, 2009-08-23.

[12] Zarrella, D. Science of Retweets, http://danzarrella.com/the-science-of-retweets-report.html, 2009.

[13] Zhao, D and Rosson, M.B. How and Why People Twitter: The Role That Micro-blogging Plays in Informal Communication at Work. Proc GROUP'09, 243-252.