Impact of Multimedia in Sina Weibo: Popularity and Life Span

Xun Zhao[1], Feida Zhu[1], Weining Qian[2], and Aoying Zhou[2]

[1] School of Information Systems, Singapore Management University
[2] Software Engineering Institute, East China Normal University

Abstract. Multimedia contents such as images and videos are widely used in social network sites nowadays. Sina Weibo, a Chinese microblogging service, is one of the first microblog platforms to incorporate multimedia content sharing features. This work provides statistical analysis on how multimedia contents are produced, consumed, and propagated in Sina Weibo. Based on 230 million tweets and 1.8 million user profiles in Sina Weibo, we study the impact of multimedia contents on the popularity of both users and tweets as well as tweet life span. Our preliminary study shows that multimedia tweets dominant pure text ones in Sina Weibo. Multimedia contents boost popularity of tweet as well as users. Users who tend to publish many multimedia tweets are also productive with text tweet. Finally, we demonstrate that tweets with multimedia contents survive longer than text tweets. Our research demonstrates the impact of multimedia in Sina Weibo with respect to how it affects the popularity, life span of tweets and the popularity of user. The results could be leveraged by social-media-based marketing and decision-making.

Keywords: Multimedia, Social network, Information diffusion, Microblog

1 Introduction

The recent years have seen social network services gaining ever-increasing popularity as a result of people's growing communication demand as well as Internet's permeation into everyone's daily life. These services have profoundly changed the way people acquire knowledge, share information and interact with one another on a societal scale.

Microblogging services, such as Twitter and Sina Weibo, allow users to publish short messages called "tweet" or "weibo" which contains no more than 140 characters. Each user may "follow" another user to receive all up-to-date messages published by that user, and get "followed" by other users to spread his messages. One can also use "@" to address a user directly. The ease of usage and succinct nature of tweets have made possible the swift propagation of news and messages in Twitter network[7].

The huge number of users, together with the staggering amount of content people generated everyday in these microblogging sites has lead researchers to analyze the syntactics and semantics underlying these social network services.

[8] points out that twitter is more of a news media than a social network. [1] points out follower count alone could not reflect the popularity of users.

Previous research on microblogging services relies mainly on textual information and social link information. However, what has as yet been largely neglected is another aspect of the microblogging data, the multimedia content, which has manifested its importance with the ever-increasing volume of the data and the profound changes it has given rise to the information diffusion throughout the network. As the saying goes — a picture is worth a thousand words. Nowadays social media users find it much more convenient and enjoyable than ever before to express their opinions by posting pictures, attaching video clips rather than just typing a message. Mobile social network application developers also introduce features to allow users to take pictures and then upload them through a simple click. Compared with text information, multimedia contents are more eye-catching and entertaining. The result is that multimedia content such as audio tracks, images and videos command viral popularity everywhere they go ranging from personal blogs, video sharing sites, to social network services. For example, according to our findings, more than 30% of tweets published in Sina Weibo contain image links. Less measurable but no less profound is the ever growing attention people paid to multimedia content, which is demonstrated by our results that, compared against tweets of pure text, tweets with multimedia content are re-tweeted by users for a much longer period of time, which we call they *survive* longer.

We focus our study in this paper on Sina Weibo, a popular, Twitter-like microblogging service platform originated from China. It features more than 300 million active users in February 2012. Besides microblogging features as those provided by Twitter, Sina Weibo has incorporated multimedia-friendly features such as attaching images as well as short url links to a tweet. Our Sina Weibo data set contains more than 1.8 million users, 230 million tweets, of which 111 million are original tweets. We show that (I) The majority of the tweets in Sina Weibo are tweets containing multimedia contents, (II) Multimedia contents such as images and short urls linked to videos are not often used simultaneously, and (III) Multimedia contents generally survive a longer period of time.

2 Related Work

On one hand, previous research on social media provide a rough tour guide of popular microblogging services. [8] uses a huge data to illustrate the user composition, trending topics et. of Twitter. [7] studies the underlying motivation of certain user activity. On the other hand, some works focused on detailed problems based textual information and social link information of the data. Various methods have been proposed to discover certain event or topic[15][17], discover the community structure [5], or trace the way information is propagated in social networks[11][6].

Another line of work analyzes multimedia contents associated with semantic geographic annotation. [2] developed methods to determine the location of a photo. This line of work is based on data that contains just multimedia contents.

To the best of the author's knowledge, this is the first work combining and comparing the textual and multimedia information in microblogging service.

3 Research Questions

Two basic elements in microblog services such as Sina Weibo are users and tweets. Users are creators and consumers of tweets. On one hand, users generate tweets by composing, publishing, or reposting tweets. On the other hand, users consume tweets by reading, reposting and replying tweets. In traditional text world, the generation and consumption process is quite straightforward. However, if we take multimedia content into consideration, would some previously identified patterns change? Specifically, we consider the following two dimensions,

1. **Tweet Generation.**
   (a) Would multimedia content influence the popularity of users?
   (b) Are users who publish more tweets also inclined to publish more tweets with multimedia content?
2. **Tweet Consumption.**
   (a) Would multimedia content influence the popularity of tweets?
   (b) Is multimedia content related to the life span of tweets?

Road Map. The following sections are organized as follows: Firstly, we give a description of our dataset in Section 4. We then analyze in Section 5 the composition of multimedia content in Sina Weibo. We explore the correlation between multimedia content and popularity. We also analyze whether users exhibit same taste for publishing text tweets and multimedia tweets. Finally, in Section 6 we illustrate the life span comparison between tweets with multimedia content and text content respectively.

4 DataSet Description

We use a corpus of data containing 230 million tweets published by 1812701 users from Jan. 2011 to Jul. 2011. In this set of tweets, 111 million are original tweets while the rest are retweets and replies. The majority of the tweets are written in Chinese.

Based on the genre of multimedia content a tweet contains, we divide tweets into the following classes.

1. Text Tweet. Text tweets are tweets which only contain text information.
2. Image Tweet. In Sina Weibo, there is a feature in each tweet indicating whether this tweet has a image link.

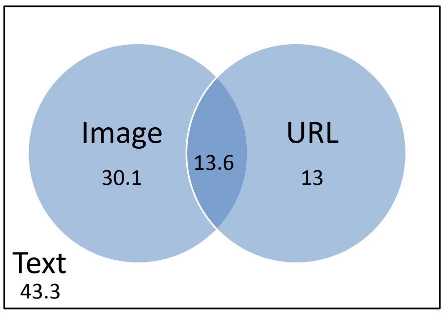
3. Url Tweet. Urls are links other than images which embed in the text body of the tweet.

Image tweet and URL tweet together forms the concept of multimedia tweet.

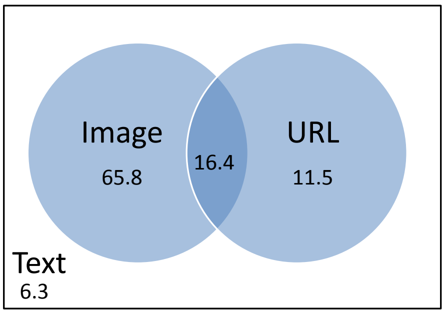
On the other hand, Sina Weibo allow users to choose whether to include a URL link specifying a homepage, favorite links or other microblog account in their profile. For ease of discussion, we categorize the set of users into 2 types, referred as URL users and NOURL users based on whether there is a URL link embedded in their profiles or not.

5 Multimedia Contents Popularity

In terms of the form of a tweet, a tweet is either an original tweet, a reply, or a retweet. Original tweets are tweets directly composed by the user and reflect the original intention of that tweet, while retweets are just reposts of original tweets and replies are commentaries about the original tweet started with a "@". Replies and retweets are widely used as measures of popularity of the original tweets [13], [8]. To study the composition of multimedia content in Sina Weibo, we distinguish between the set of General Tweet and Popular Tweet. General Tweet consist of all the original tweets in our dataset and Popular tweets are a subset of General tweets which receive a considerable amount of retweets. [1] has reported that popular tweets are more likely to be posted by celebrities and news medias. [18] has reported the topics of popular tweets are different from ordinary tweet. [16] finds out that the trends in Sina Weibo are created due to the retweet of multimedia content such as jokes, images and videos. Our analyses further support this point.



(a) General Tweet Composition



(b) Popular Tweet Composition

Fig. 1. Venn Diagram for Composition of Multimedia Tweet

5.1 Original Tweets Content Composition

127 million out of 230 million tweets in our dataset are replies or retweets. Replies and retweets are comments and replicates of original tweets. They can be used as measures for popularity of original tweets[1], but they do not have any content value. For original tweets, which are not replies nor retweets, we divide them into 3 categories, namely, Text Tweet, Image Tweet and URL Tweet as previously categorized. We also select another group original tweet which received more than 1,000 retweets for comparison. We call this set of tweet Popular Tweet. Figure 1 shows Multimedia content (Image and URL) composite more than 50% in both setting. In more detail, Image Tweets dominate in general tweet composition, with more than 40%, the dominance is more profound in popular tweet setting with text tweet only composite 6.3% in popular tweets. This shows while text tweets do exist in a considerable amount, the majority of trending tweets in Sina Weibo are multimedia content tweets. Interestingly, we also see a no small overlap between image tweets and URL tweets, which indicate the usage of multimedia is integrative and simultaneous.

A similar approach is to use the number of replies to define the popularity of a tweet. To our regret, our data does not contain reply information. Thus we do not provide popularity analysis based on replies in this article.

5.2 Tweet and User Popularity

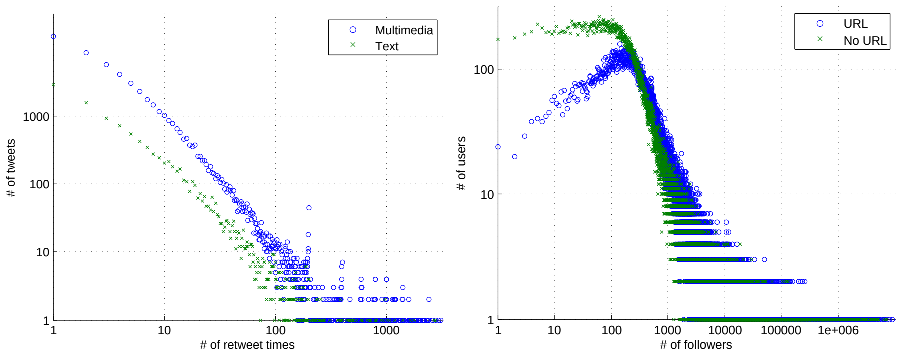
To understand the interplay between popularity and multimedia content, we need to examine the popularity each tweet and user dissents and the difference between multimedia content and plain text. To measure the popularity of tweet and user, we follow the convention in [1] and use retweet times as a measure of tweet popularity, and follower count for user popularity.

Figure 2(a) displays tweet popularity distribution of 1,000,000 randomly selected tweets. The overall distribution approximately fits a power law pattern[4] with most of tweets receive very few retweets and only a few tweets receive large number of retweets. The number of tweets from different popularity level differs by orders of magnitude. Interestingly, we also observe a long tail in both multimedia setting and text setting when $retweettimes > 100$. This abnormal pattern indicates the number of very popular tweet is larger than power law distribution suggests, reflecting that very popular tweets do exist in a considerable amount. This finding has important implications for microblog based marketing. Marketers would get a great pay off by aiming at those top popular tweets.

The proportion of multimedia tweets in these 1 million tweets is 61.8%, which is consistent with our previous composition analysis in general setting. With retweet number set, the number of multimedia tweet is larger than text tweet. While with tweet number set, retweet times of multimedia tweet is also larger. This reflects that multimedia tweets are more popular than text tweet in terms of absolute number and retweet times.

Sina Weibo allow users to include another type of multimedia content right into their profile. In their profile, a user could put a url-specified homepage link, blog site or other microblog account. Based on whether a user puts such url links in their profile, we divide users into two groups. For simplicity, we use URL to refer to the set of users who have such information, and No URL for those who do not.

Follower count could be used as a measure for user popularity [1]. Figure 2 (b) also shows a power law pattern when $200 < follwercount < 1000$ for both set of users, as the number of users decreases exponentially with follower count increase. We also observe a long tail when $followercount > 1000$, indicating the number of very popular users is more than the power law pattern suggests. For URL distribtuion, we find a global maximum at $follower = 200$. While before URL distribution reach its peak, the number of NO URL users is always bigger than URL users. We conjecture that this may result from the fact that URL users tend to engage more effort in maintaining their Weibo account as well as interacting with their friends, making the number of inactive(less followers) users less than NO URL users.



(a) Tweet Popularity Distribution (b) User Popularity Distribution

Fig. 2. Tweet and User Popularity Distribution

5.3 Comparing User Activeness

For multimedia content lovers, are they also craving in posting a lot of text tweets? Specifically, are users who publish most multimedia tweets also the ones who publish most text tweets? The amount of tweet a user posts can be used as an indicator of user activeness[1]. We get the number of text tweets and number of multimedia tweets for each user in the previous setting. Rather than directly compare the number of text tweets and the number of multimedia tweets, we use the relative order of user ranks based on tweet quantity and multimedia quantity as a measure of difference. We first sort users by those two measures, so the rank 1 user in tweet quantity indicates the most active publisher. Increased ranks imply less active publishers. Users with the same number of tweet would receive the average rank amongst them. Once each user receives a rank from these two measures, we could compare their rank difference. We use Spearman's rank correlation coefficient[12]

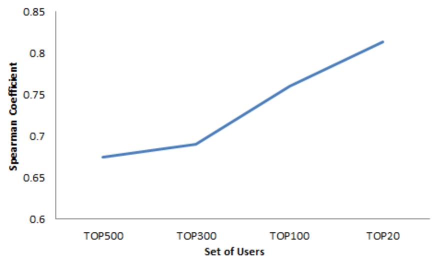
$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N^3 - N} \tag{1}$$

as a measure of the strength of association between two rank sets, where x_i and y_i are ranks of users based on two measures in a dataset of N users. The coefficient assesses how well a monotonic function could describe the relationship between two variables, without making any other assumptions about the particular nature of the relationship between the variables. The closer ρ is to $+1$ or -1, the stronger the correlation. A perfect positive correlation is $+1$ and a perfect negative correlation is -1.

The results in Fig.3 show a moderate strong correlation(above 0.6) between ranks of multimedia tweet quantity and text tweet quantity for all pairs. However, if we narrow our focus on top 500 users, those who rank top 500 in tweet quantity, the correlation becomes stronger. Further narrowing on even top users lead to even higher correlation, indicating users who publish most tweets also publish most multimedia tweets, especially for most active users.

Set	ρ
All	0.638
Top 500	0.675
Top 300	0.690
Top 100	0.760
Top 20	0.814

(a) Spearman Correlation Samples



(b) Spearman Correlation Trend

Fig. 3. Spearman Correlation of User Post Activeness

6 Life Span Analysis

Many factors, such as user popularity and topic of tweet[10] could affect the life span of a tweet. Previous studies [7][8] have reported that messages in microblogging services such as Twitter spread and disappear rather fast. [9] reported that instead of a social media, Twitter is indeed a broadcast medium with virtually all retweets happens within the first hour after the original tweet. Figure 4(a) shows how retweet times changes for a typical tweet as time passes in [9]. It quickly receives a lot of retweets after its birth and slowly lose its attention.

Interestingly, in our Sina Weibo data, we find that some of the tweets remain viral and repeatedly get reposted for a long period of time. To get the temporal effect of multimedia contents, we set up the following experiment: We first select all trending tweets, including retweets and original tweets, which get at least 500 retweets in July, 2011. Then we filter out retweets and get the original tweet id of these trending tweets. Finally, we go to previous months and check the publish time of these original trending tweets. We only track 6 months backwards, which start from January to June, for original tweets found prior to January is too small for statistical analysis.

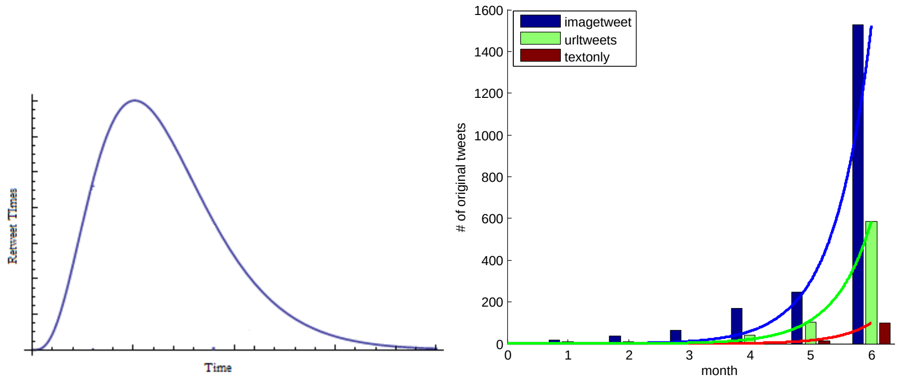
For comparison, we also separate the trending tweets into three categories: Text Tweet; Image Tweet; URL Tweet. Figure 4 shows the bar plot of how many original tweets within each category are found in each month from January to June.

According to [11], life span of memes, or new topics, follows exponential decay. In this article, we follow this convention and model the life span of tweet as the form

$$N(t) = N_0 e^{-bt} \tag{2}$$

where $N(t)$ is the quantity at time t, N_0 is the initial quantity and b the decay rate.

Based on this assumption, we use Formula 2 to fit the data observed in Figure 4(b). We use non linear least square method to fit the data. Table 1 shows the result coefficient of the fitting, where $\tau = \frac{1}{b}$ is defined as the mean life span of tweets.



(a) Gamma distribution representing a typical tweet life span in Twitter

(b) Number of original tweets found in each month and fitting result

Fig. 4. Tweet and User Popularity Distribution

The amount of original tweets in Figure 4(b) shows all three groups drop exponentially from June to January. The amount of image tweets are always dominant in each month followed by URL tweets, further suggesting multimedia content's power of attracting retweets over text. The decrease rate, however, is a bit different among three groups. As in Table 1, text tweets have the largest decay rate, followed by URL and image tweets, which implies image tweets have the longest life span, followed by URL tweets and text tweets.

In Table 1, there is a significant gap between life span of Text Tweet and the other two multimedia groups, while the difference between Image Tweet and URL Tweet is marginal. This shows a fundamental difference of content virality as well as popularity between multimedia tweets and text tweets. This is because the rich information and eye catching nature makes multimedia tweets more viral than text tweets, thus enabling them to spawn a longer period of time after they first get published. For comparison in the two multimedia group, image tweets show a slightly longer life span than URL tweets. We conjecture that this is because pictures are directly embedded in the tweet, which gives users a direct visualization, while URLs are more often appeared as links, and content illustration is dependent on the text information rather than multimedia itself.

Table 1. Decay rate coefficient with error range

Category/Coefficient	N_0	b	τ
Text	0.001678(-0.004541,0.007897)	1.831(1.211,2.451)	0.546
Image	0.08929(-0.231,0.4096)	1.624(1.022,2.226)	0.616
URL	0.02766(-0.02104,0.07636)	1.660(1.365,1.955)	0.602

The error range of Text group is larger than the other two groups. We conjecture that this is caused by the small amount of data in text tweet. Only a handful of text tweets are found in the beginning months of the year.

In order to get a larger sample size, we also set different retweet popularity threshold(100, 200, 300, etc.) in this experiment. In all of these attempts, the program would not finish running because of large sample size. Our findings point out that multimedia content have a longer life span than traditional text messages.

7 Conclusion

In this paper, we show multimedia tweets composite a large proportion in Sina Weibo. Moreover, we demonstrate multimedia contents influence the popularity of tweet and user by boosting the retweet times of a tweet and the follower number of a user. The number of highly popular tweets exists in a larger scale than power law pattern suggests. Multimedia contents help to promote retweets and follower account of user. Users who publish large number of text tweets are the ones who publish a lot of multimedia tweets. Finally, we study the correlation between multimedia contents and tweet life span. Multimedia tweets such as image tweets and URL tweets have a longer life span than text tweet.

8 Acknowledgement

This work is partially supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. This work is also supported by the China National Basic Research (973 Program) under grant number 2010CB731402, and China National High-tech R&D Program (863 Program) under grant number 2012AA011003.

References

1. M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.
2. D.J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770. ACM, 2009.
3. A. Cui, M. Zhang, Y. Liu, and S. Ma. Are the urls really popular in microblog messages? In *Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on*, pages 1–5. IEEE, 2011.
4. D. Easley and J. Kleinberg. *Networks, crowds, and markets.* Cambridge Univ Press, 2010.
5. J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 813–822. ACM, 2010.
6. M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):21, 2012.
7. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.
8. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
9. F. Lardinois. The short lifespan of a tweet: Retweets only happen within the first hour. In *http://www.readwriteweb.com*.
10. K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.
11. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

12. C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.
13. D. Wang, Z. Li, K. Salamatian, and G. Xie. The pattern of information diffusion in microblog. In *Proceedings of The ACM CoNEXT Student Workshop*. ACM, 2011.
14. Sina Weibo. http://www.weibo.com.
15. J. Weng and B.S. Lee. Event detection in twitter. In *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM)*, 2011.
16. L. Yu, S. Asur, and B.A. Huberman. What trends in chinese social media. *arXiv preprint arXiv:1107.3522*, 2011.
17. W. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*, pages 338–349, 2011.
18. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.P. LIM, and X. Li. Topical keyphrase extraction from twitter. In *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics*, 2011.